UCLA MAE CAPSTONE

USING STATISTICAL AND MACHINE LEARNING TECHNIQUES TO PREDICT LOS ANGELES REAL ESTATE PRICES

Gary Vartanian

Abstract

Real estate home prices are tricky to properly evaluate. This is because there is no objective way to value your home. It is combination of different factors such as: location, home features, and ultimately how much someone else is willing to pay for it. With the abundance of data science techniques, I wanted to see if there is way to accurately create a model that can help explain the importance and weights of each variable as well as provide accurate results. Different Machine Learning and Statistical Learning Techniques such as: Linear Regression, Stepwise Regression, Lasso & Ridge Regression, Regression Tree, Random Forest, and Neural Network are applied to the benchmark real estate standard of comparables (proxied via K-nearest neighbors) for home prices in Los Angeles Area. The end result is to have a model that can beat the current real estate standard in both accuracy and interpretability.

Thanks and Acknowledgement

I would like to give a special shoutout to the supervisors on this paper who believed in me and let me do this under them. Thank you Randall R. Rojas, Patrick Convery, and Melody Huang. You guys are awesome! I would also like to thank the Masters in Applied Economics program for a great year and giving me the opportunity making alot of new friends, learning new material, and helping pick up the skills to be an economist. Lastly, I want to thank MAE director, Aaron Tornell, for inspiring me to go into economics. One last shoutout goes to my sister, Lara Vartanian, for assisting me and pushing me in this capstone project.

Capstone Introduction

Real Estate has been of particular interest to me for a while. It is one of the oldest and biggest assets yet when asked how to determine its value, no one person can provide the same answer. Granted there are similar valuation techniques that people use worldwide. It was this interest that got me originally started.

Current Methods for Valuating Real Estate

The current mainstream methods to evluate homes are as follows:

1) Appraisal
2) Comparables
3) Automated Valuation Models

Appraiser

An appraiser is an industrial professional that values a property and provides an estimate based off their experience and expertise. An apprasial is the appraiser appraising your property and providing you an estimate. However, different appraisals will provide different answers. This is because expertise is relatively subjective. One can say that this house is worth $500,000 while another can say its worth $500. Realistically, you don't expect that much deviation in the real world. For example, how many professionals would say a home in Beverly would cost $500? If you hire professionals, then theoretically they should all come up with similair estimates.

This process fascinates me as it almost exclusively relies on people. It is grounded in opinion, heurisitcs, and expertise and not really much grounded model. I want to go about providing more of an explanation? How do we know if they are right or wrong about the price? Could I have sold my home for less or more? If I was to attach an additional bedroom or bathroom or garage, would that change anything? How would the price change if it was in another location? These are questions that would be hard for the expert to answer right off the bat and motivated me to look into other methods.

Comparables

The other method is comparables. This is the idea that if you want to sell your home, you look at the recently sold homes that closely matches the criteria of your home. For example, if I want to sell a 2 bed 2 bathroom apartment in Westwood. I would look at other recently sold 2 bedroom, 2 bathroom apartments in Westwood. It stands to reason that if someone bought that previously for that price, they would be willing to pay for it again.

This model has been the real estate golden standard for many years and the way many people value the price of their home.

Automated Valuation Models

An automated valuation model, is as the name implies a valuation model that is automated. It is essetnially a mathematica model that guestimates the value of a home. It picked up in the late 1990's and is used currently by many different players in the real estate industry. For example, Zillow and Redfin (two big players in the industry) use their own proprietary models to help evaluate the home. As you can imagine it has many advantages: it saves time, money, and resources, by simply reducing it down to an algorithmn that you can plug and apply to any household. This is much more cost effect than highering someone to evaluate door to door. AVM's remove the human element from the valauation process and rely on computer automation so as to remove human bias and subjectivity.

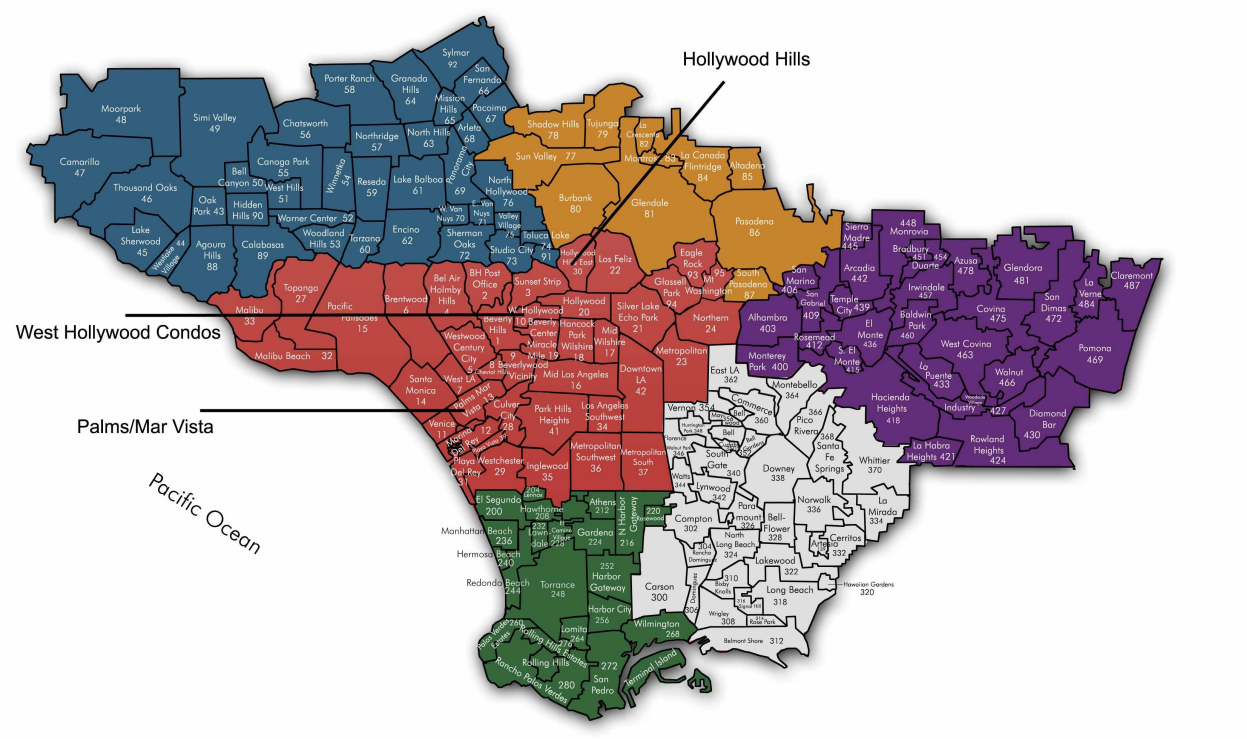


Figure 1: "Breakdown of Different LA Areas"

The problem with automating prices with a model is that you ignore the current condition of the property. On many databases, it is extremely hard to quantify the condition of the property therefore it cannot be counted as an input and goes unnoticed. Zillow might say that the haunted house down the street is worth a million dollars based off the sqft, bedroom, bathroom, and location, but if its barren, bug infested, and eroded, the model would totally miss that.

Although, it has its shortcomings the main objective would be to see that if an accurate Automated Valuation Model can be constructed to help predict home prices.

Data Explanation

The data I have obtained is from the Los Angeles: Multiple Listing Service (MLS) and provides the most recently home prices. MLS is the primary service and database for real estate brokers and fastest information on real time sales.

The data chosen is the Los Angeles Real Estate homes that were sold in the last year in 2017. Los Angeles is a big area and its subdivided into different regions within in (See Figure 1). Around 2500 homes were sold last year. As you may be aware you have some areas that are hot (lots of buys) and some that are cold (low number of buys). With MLS, all real estate transaction are recorded for the sake of regional simplicity. I decided to choose the top 4 areas with the most transactions: Los Angeles Southwest, Metropolitan South, Metropolitan Southwest, and Mid Los Angeles. These 4 areas had around 1400 of the transactions. Because of their hotstreak value (interest) and because of the abundance of data points in that area, these will be the main ones focused on in the regression model. Furthermore, each one of those areas have 200+ transactions in the dataset, helping provide a stable model. Other areas on the original list range from 18-30 data points, which I felt did not provide strong support for regression techniques.

The main variables are the following:

1. **Type**

- This indicates whether apartment or home. In this dataset and study, the primary concern is for houses, therefore the focus on the label (SFR - Single Family Residential). (factors)

2. **MLS.area**

- This variables talks about the MLS area that the sold home corresponds to. There are four variables: Los Angeles Southwest, Metropolitan South, Metropolitan Southwest, and Mid Los Angeles. These are all counted as factors. (factors)

3. **Housing Price**

- Denoted as h.price in our dataset. this is the primary variable of interest and the price that the home was last sold at. (numeric)

4. **Sqft**

- This variable is the number of square feet of the home. (numeric)

5. **Bedroom**

- This variable states the number of bedrooms within the home (numeric)

6. **Bathrooms**

- This is broken down into four categories: full bath, 3/4 of a bath, half bath, and 1/4 bath. The math is simple: Each utility is counted as one-quarter, so you add and deduct a quarter for each one, as the case may be. Therefore, a bathroom with a sink, toilet, and shower is considered a three-quarter bath. A bathroom with just a sink and a toilet is a half-bath...

7. **Full Bath**

- This states the number of full bathrooms within a home. A full bathroom is made up of four parts: a sink, a shower, a bathtub, and a toilet. (numeric)

8. **Three quarters of a bath**

- This states the number of bathrooms that satisfy the condition of 3/4 of a bath. Again, a bathroom with a sink, toilet, and shower is considered a three-quarter bath. (numeric)

9. **Half Bath**

- Number of rooms in the house that meet condition of half bath. A bathroom with just a sink and a toilet is a half-bath. (numeric)

10. **Quarter Bath**

- Number of rooms in house that meet condition of half bath (just sink or just toilet). (numeric)
- Quarter bath was originally included but was later removed in due to the fact that there were less than 13 homes that contained a quarter bath. This could not adequately support in any sort of regression model. (numeric)

11. **Year Built**

- As the name implies it corresponds to the year that the house was built. Unfortunately, there is no information on if there was renovation or not in the home. (numeric)

12. **Pool**

- This is a binary variable indicating if there is a pool within the home. (numeric)

13. **Garage Space**

- This is a binary variable indicating if the home had a garage.

14. **Longitude**
    - This corresponds to the longitude coordinates that the home is on.
15. **Latitude**
    - This corresponds to the latitude of the home's location.

In total, you have 14 predictors (independent variables) and 1 dependent variable (home price) that want to predict.

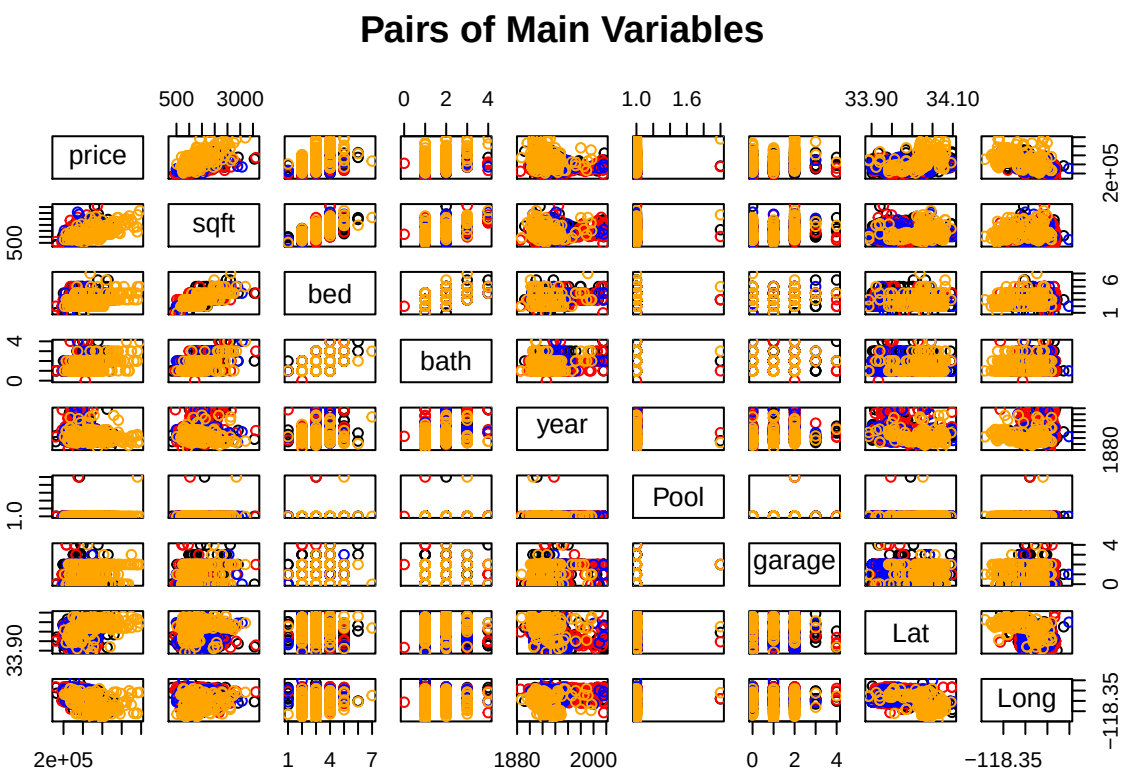
Exploratory Data Analysis

In the following section, exploration of the data will begin to uncover different relationships, details, and underlying distribution of our variables.

The visualization will be approached by the following:

1. Pairs
2. Bar Plot and Density Plot
3. Comparison of Distributions
4. Corrplot
5. Boxplots and Violinplots
6. Variable Exploration
7. Mapping using Google Maps
8. Clustering with K-means Clustering
9. Bi-Plot
10. Plotting of Regressions

Pairs



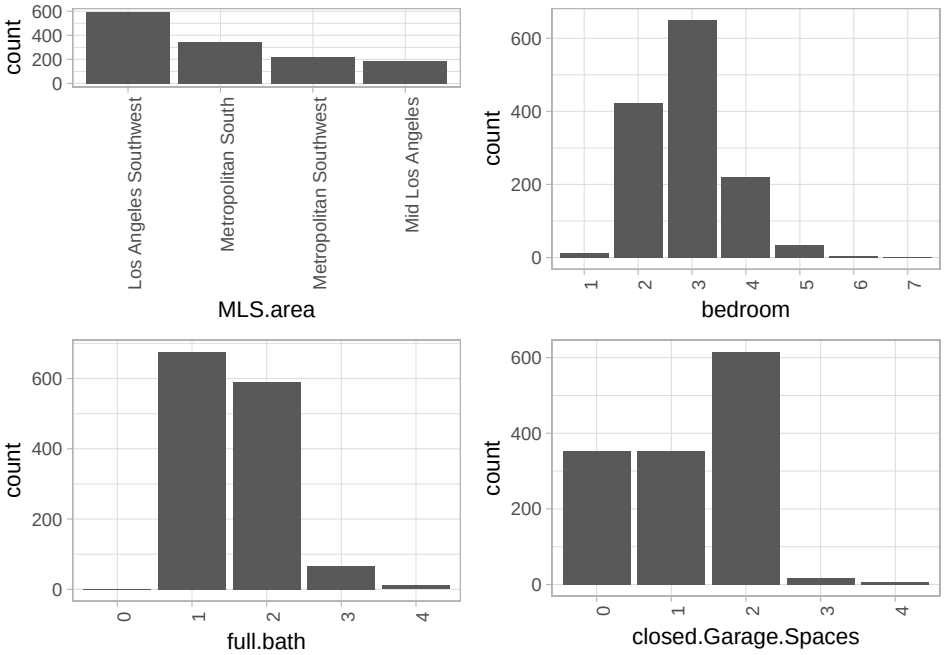
The plot is a pairs plot of all the variables. The color is representative of the area that they come from. Black, Red, Blue, Orange correspond to Los Angeles Southwest, Metropoitan South, Metropolitan SouthWest, and Mid Los Angeles. It is supppose to provide a quick glance at what is going.

Firstly, we notice a few things: there is a linear relationship between housing price and sqft. Pool is a tiny binary variable with not many observations. Relationship between bedroom, bathroom, and sqft, which does make sense cause the bigger the house the bigger the following.

More interestingly, I want to look at the different pairs we have going on. There doesn't seem to be much color differntial between all the pairs. The main distinguishment is from the variable year built and longitutde and latitude. Longitutde and Latitude intuitively make sense because they correspond to the area on the map for these locations. Though not far apart from each other, they do help show the difference in areas. On the other hand, it is extremely interesting to see year built variable stand out with distinguishing areas. From the insights above, we can see that developement in the different areas started in different years and that Mid Los Angeles (orange) started in the early years and construction slowed down.

Bar Plots and Density Plots

Bar Plots



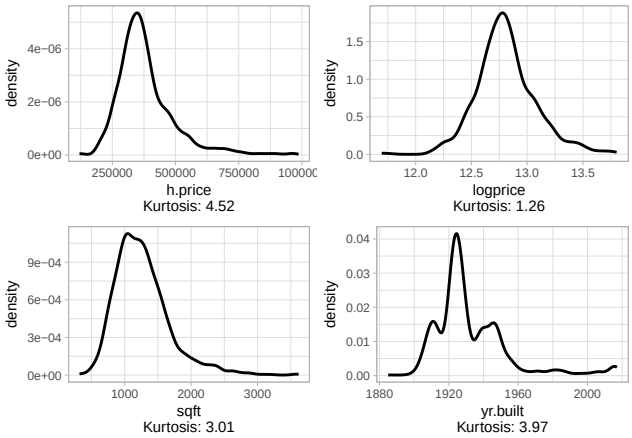
The first visualization breaks down the frequency for the different regions, bedroom, full bathrooms, and number of garage spaces. From the top left we see the number of single family residential homes that we have from each area in our data set. Most of the data set is Los Angeles Southwest (n $\approx$ 600), while Metropolitan South (n $\approx$ 300) and Southwest (n $\approx$ 250). Lastly, Mid Los Angeles with $\approx$ 200. Looking to upper right we see that the average number of bedrooms is around 3 with a nice bell curve. In regards to bathrooms, most of the data is skewed to 1 or 2. Lastly, for garage spaces nearly all the data indicates that most people only have up to 2 garage spaces.

Density Plots

These are a quick visualizations of the density of our data. In regards to price, we can see a normalish distribution but disturbed by fat tails. Alternatively, by taking a log of the price, we normalize it more,

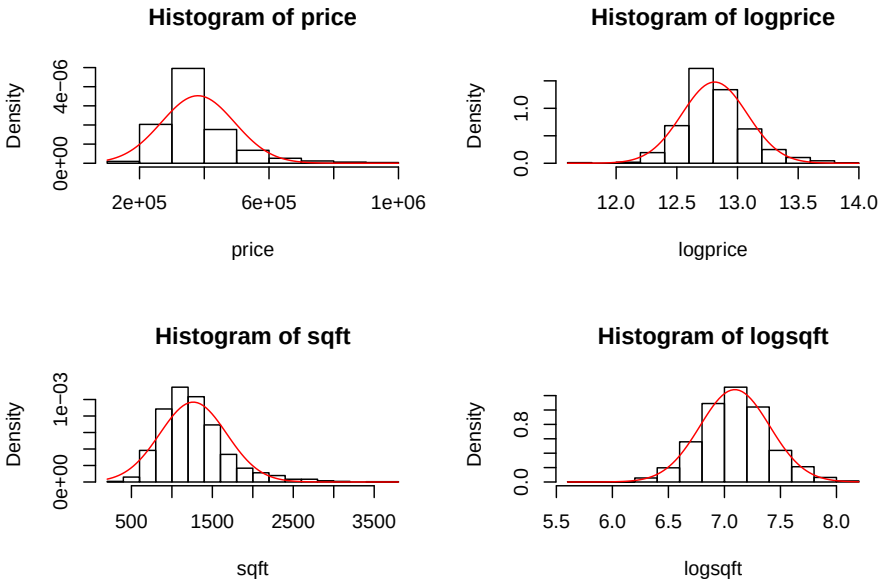
but still not exactly conforming. Sqft follows the same path with most of the homes being between 900 to 2000 sqft. Year Built is rather interesting as we some some changes. We see a build up all the way up to 1930's and then following great depression a crash in the number of homes. We then see a gradual slip in construction of this area.

Traditionally, kurtosis has been explained in terms of the central peak. Higher values indicate a higher, sharper peak; lower values indicate a lower, less distinct peak. Balanda and MacGillivray (1988) also mention the tails: increasing kurtosis is associated with the "movement of probability mass from the shoulders of a distribution into its center and tails." However, Peter Westfall (2014) defines that "higher kurtosis means more of the variance is the result of infrequent extreme deviations, as opposed to frequent modestly sized deviations." In other words, it's the tails that mostly account for kurtosis, not the central peak.



A normal distribution has kurtosis exactly 3 (excess kurtosis exactly 0). By examining the kurtosis above, we see that the price is hindered by high kurtosis (extreme outliers) and this can be helped by log adjusting it. The sqft has a kurtosis of that of a normal distribution, yet does not precisely follow the normal. Theoretically, the 2 main important variables are housing price and sqft, therefore we would want to see if they both follow a normal distribution. As a result, I've plotted housing prices and sqft and their log equivalent and overlayed a normal distribution on it for comparison.

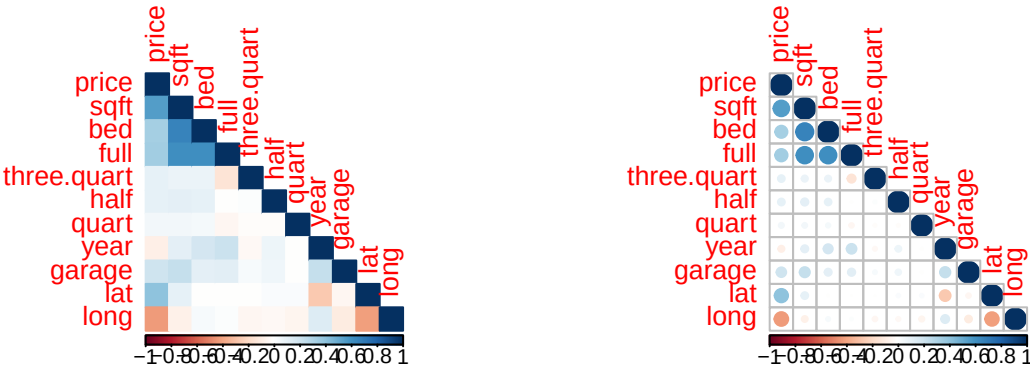
Comparison to normal



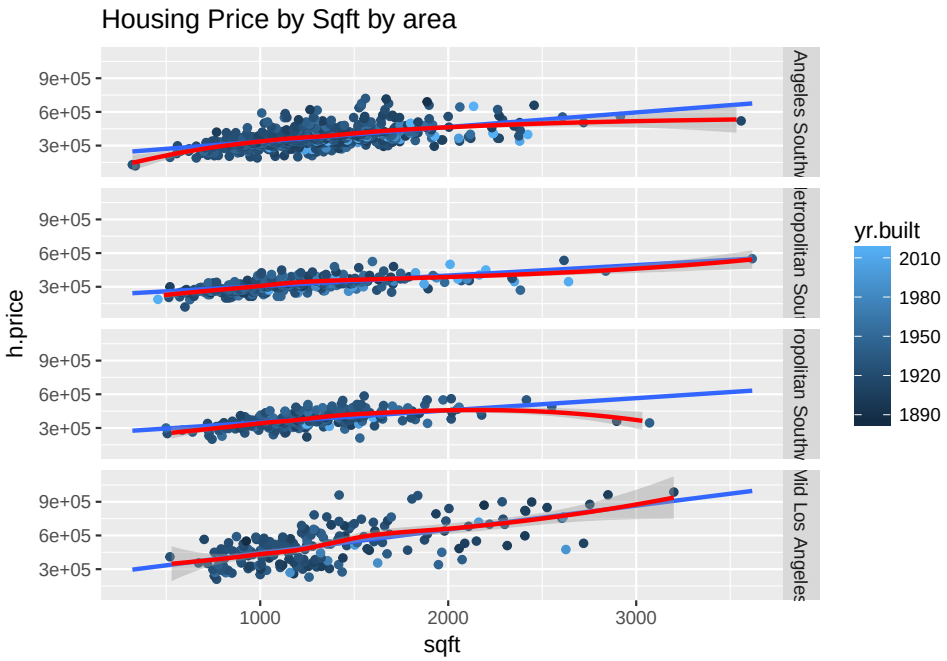
The visualizations above show a histogram of the 2 most important variables and a normal distribution overlayed with them. Right off the bad we see that the variables in the original state do not match normal, but upon transformation closely match. However, to be certain you'd have to run a Jarque Bera test on the variables. Sadly, it reported that all did not follow the normal distribution. With that done, inspection of the correlation would commence.

Corrplot

This is a correlation plot of all of the variables. The main thing you want to observe is the variables that have a strong correlation with price. So far looking down the first column you can see that the main factors that correlate with price are sqft and Longitude. Other features of the house such as bedroom, bathroom, and latitude also relate. These all make sense as the factors to consider when your buying a home. My belief of why sqft has a stronger correlation has to do with the interaction between that and the areas. I theorize that the price per sqft changes in different areas due to zoning permits, median income, and other features as a result highly correlating with price. For bedroom and bathroom, although these are important features, I cannot imagine them varying that much in price in different areas. My personal hypothesis is that it would be of constant value throughout. Lastly, latitude and longitude correlate with price and this is to be expected as the most important variable in real estate is: location, location, location. Just because there is a low linear correlation doesn't necessarily mean that there is no insights provided. I am confident that there is some sort of non-linear impact of these variables.

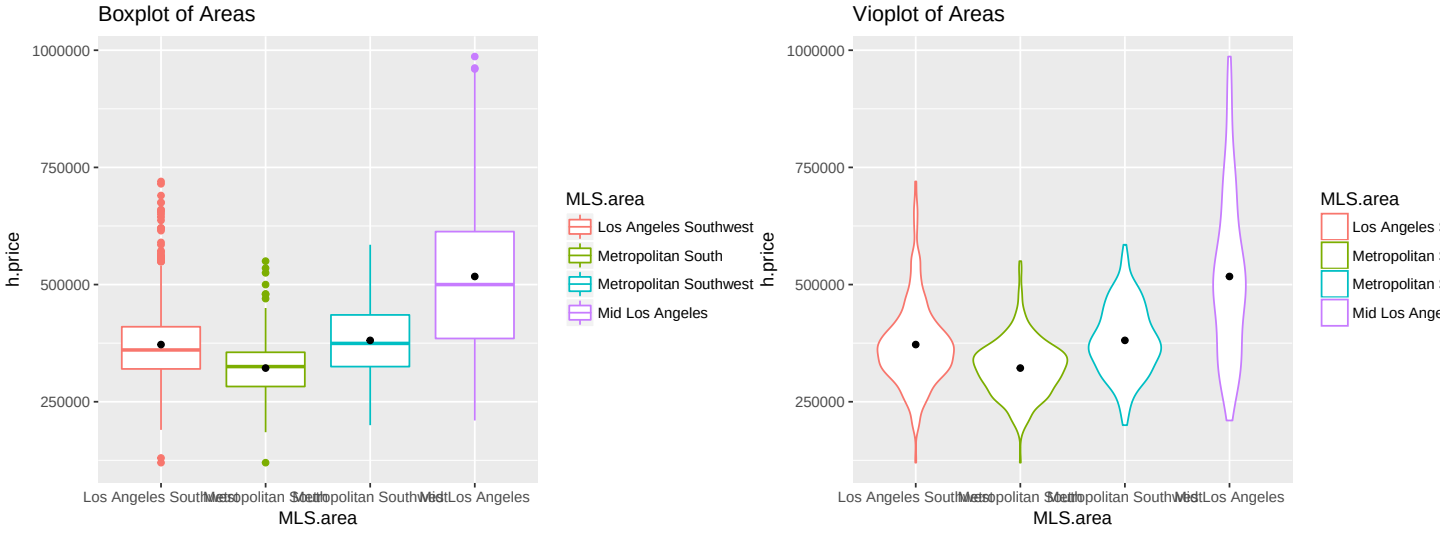


Variables by Area



The above graph has the housing price by square feet per area. Furthermore, it is colored by the year that the house is built. In regards to the colors, something that stands out is the newer houses are more to the left hand side indicating that they have bigger square footage. Overall, you can notice a strong linear trend between the different variables. As a result, I thought it was fitting to run a linear regression and compare it to a smoothes curve. For the most part, the smoother closely follows the linear model, especially in the middle. Indicating a good sign that the linear model captures the points well. Unfortunatey, the smoother breaks towards the end due to the lack of points. From the image it looks like linear models would do great for the home.

Boxplot and Vioplot



With the different areas plotted above, it might be best to examine them through boxplot and violin plot. With the boxplot, it is noticeable that the Los Angeles Southwest and Metropolitan South have a lot of

outlier that above the mean. In particular for LASW this can be seen by the dot (mean being above the line in middle of box). Metropalitan Southwest does not have any outlier while Mid Los Angeles seems to have 2 extreme outliers. The violin plot shows similair information as the boxplot but with the density of the plotted "creating the violins". It is rather interesting to see for the first two areas, the price density is highest around the middle. For metropolis southwest you see a spike a little further up you go and gradual diminishment. As for Mid Los Angeles, you have a lot of variance going on.

Basic Summary Statistics

With that these are some basic summary statistics of our data.

```
## [1] "This describes the prices by area"
```

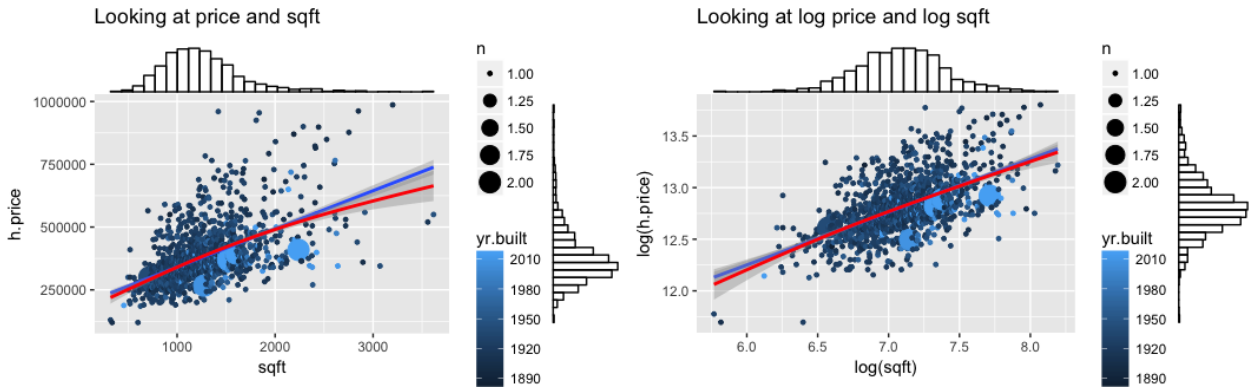
```
## R8[, c("MLS.area")]: Los Angeles Southwest
##    vars   n     mean       sd median  trimmed     mad    min    max range
## X1    1 596 371996.4 88400.35 360500 365152.7 67458.3 120000 720000 6e+05
##    skew kurtosis      se
## X1 0.91     1.74 3621.02
## ------------------------------------------------------------
## R8[, c("MLS.area")]: Metropolitan South
##    vars   n     mean       sd median  trimmed     mad    min    max  range
## X1    1 347 321852.9 59710.42 325000 320578.2 53373.6 120000 550000 430000
##    skew kurtosis      se
## X1 0.34     1.06 3205.42
## ------------------------------------------------------------
## R8[, c("MLS.area")]: Metropolitan Southwest
##    vars   n     mean       sd median  trimmed      mad   min    max  range
## X1    1 216 380928.3 73956.73 374450 379679.2 76428.03 2e+05 585000 385000
##    skew kurtosis      se
## X1 0.18    -0.29 5032.12
## ------------------------------------------------------------
## R8[, c("MLS.area")]: Mid Los Angeles
##    vars   n     mean       sd median  trimmed    mad    min    max  range
## X1    1 185 517280.3 171315.5  5e+05 505193.3 170499 210000 986500 776500
##    skew kurtosis       se
## X1 0.59    -0.13 12595.36
```

```
## [1] "This describes the prices in total"
```

```
##    vars    n     mean     sd median  trimmed      mad    min    max  range
## X1    1 1344 380483.8 113028 359900 366690.8 81394.74 120000 986500 866500
##    skew kurtosis      se
## X1 1.67     4.52 3083.09
```

Linear Model with All Variables

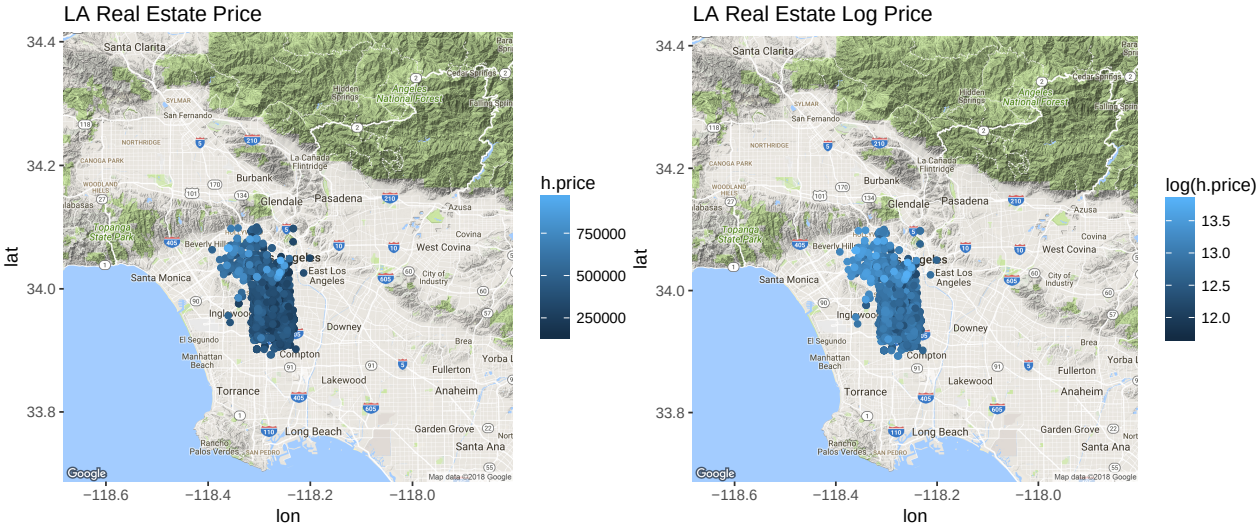
To continue off the correlation and previous exploratory analysis, it wanted to see if there is an overall linear relationship between the variables. Therefore, I ran an overarching linear regression and loess smoother Here is provided the linear graph, lm model, smoothes curve, distribution of price and sqft. Alternatively, I would like to examine the log of these 2 variables seeing how transformation would make a difference.



From the linear model we see that there is quite a big variance when it comes to the price. As the sqft increases the price of the home can vary. My suspicion falls under the different categories of areas that this can be in and how all of them are jumpled together. For example, a sqft in Beverly Hills might be worth more than a sqft in Inglewood and that is why I feel that the linear regression deviates from the loess smoother in that case.

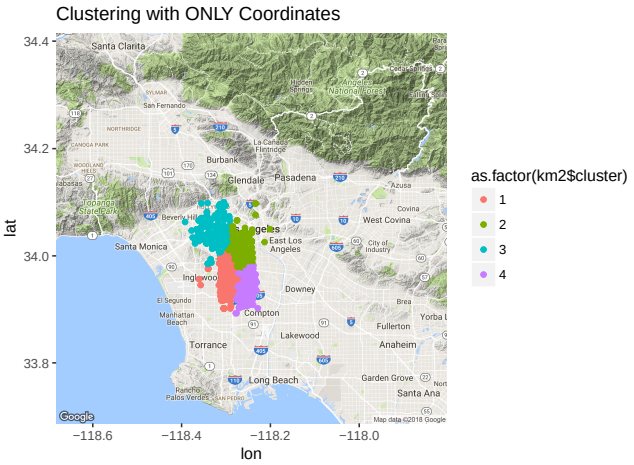
I prefer looking at the log graph graph better cause it helps normalize the variables that we wanted. In addition, it provides a strong linear fit. Examining the log of these variables, you can see constant variance over time and signs of homoscedacity.

Visualization on Map

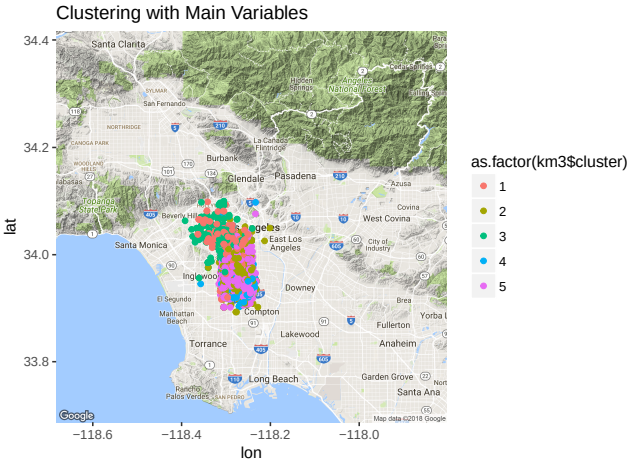


The visualization in the above plots the most recently sold single family homes within the Los Angeles market and provides visualization in regards to the price. In the first graph you can see that Latitude and Longtitude have an impact to play. Generally, the homes further north are priced high while the ones south are from the lower end. Also, you notice more lower priced homes in the right then in the left. These attributes are highlighted more through the log Price visualization.

Clustering on Map



To partition these houses into groups, I decided to use k - means clustering algortihmn to see the different divides. First, I did it with only using the coordinates. This should be very straightforward into dividing them into 4 quadrants. This is mainly to highlight the second graph which includes clustering based off all the main and important variables.



Interestingly enough, they both differ quite a bit. The first one is very straightforward and plots in 4 quadrants as expected, but the second one which is done with house price, sqft, bedroom, bathroom, year built, Longitude and Latitude are giving quite an interesting result. I've went ahead and printed the centers of these clusters to provide some more insights.

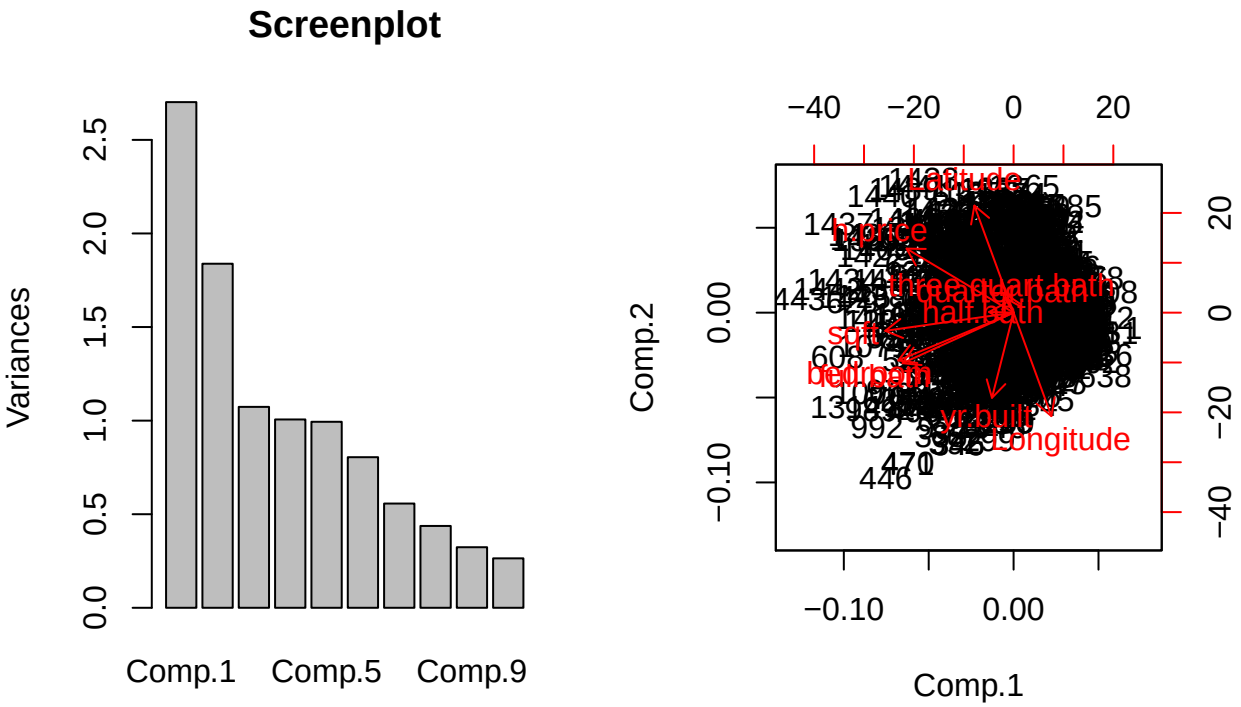
h.price	sqft	bedroom	full.bath	yr.built	Longitude	Latitude
1.9485219	1.8377191	1.2273138	1.2188767	-0.5140380	-0.6902258	0.8983530
-0.6707868	-0.7430599	-0.7188435	-0.8361127	-0.1443725	0.3268270	-0.4016901
0.5524625	-0.1374800	-0.3607906	-0.4028661	-0.2615634	-1.0811751	1.2957087
-0.0898059	0.7628910	1.1161878	1.0640641	2.4687523	0.5309646	-0.5221978
-0.0397808	0.2399409	0.4131639	0.5952852	-0.2241028	0.2822410	-0.3783253

Keep in mind that these are standardized scores. It seems to me that cluster 1 (red) is the high priced, high sqft, big bedroom houses that are older. Cluster 2 (gold) would be the cheapest area with the smallest sqft and bedroom bathroom area. Cluster 3 (green) would represent the most normal housings in regards to features and amneties. This can be see by the sqft, bedroom, bath close to zero. However they are more

expensive due to their neighborhood. Cluster 4 (blue) would represent the newest houses (we can see this by the price similair to mean and big sqft which we noticed in our eda earlier). Group 5 represents smaller homes that are priced nearly the same. These are my rough thoughts on the clustering going on.

PCA

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
## Standard deviation     1.6436782 1.3558262 1.0360357 1.0031054 0.99689951
## Proportion of Variance 0.2701678 0.1838265 0.1073370 0.1006220 0.09938086
## Cumulative Proportion  0.2701678 0.4539943 0.5613313 0.6619533 0.76133418
##                            Comp.6     Comp.7     Comp.8     Comp.9
## Standard deviation     0.89699915 0.74630042 0.66130666 0.56886502
## Proportion of Variance 0.08046075 0.05569643 0.04373265 0.03236074
## Cumulative Proportion  0.84179493 0.89749136 0.94122401 0.97358475
##                           Comp.10
## Standard deviation     0.51395765
## Proportion of Variance 0.02641525
## Cumulative Proportion  1.00000000
```



As used in Principal Component Analysis, the axes of a biplot are a pair of principal components. These axes are drawn in black and are labled PC1, PC2, etc.

A biplot uses points to represent the scores of the observations on the principal components, and it uses vectors to represent the coefficients of the variables on the principal components.

Points that are close together correspond to observations that have similar scores on the components displayed in the plot. To the extent that these components fit the data well, the points also correspond to observations that have similar values on the variables.

A vector points in the direction which is most like the variable represented by the vector. This is the direction which has the highest squared multiple correlation with the principal components. The length of the vector is

proportional to the squared multiple correlation between the fitted values for the variable and the variable itself.

vectors that point in the same direction correspond to variables that have similar response profiles, and can be interpreted as having similar meaning in the context set by the data.

In our case, we examine the variables. We see that the fractional bathrooms do not add much explanatory power cause they do not contain much variation. However, the first principle component does quite well in capturing features of the house such as sqft, full bedroom and bathroom. This is indicated by its high placement to the left. The second principal component weighs latitude, longitutde, and price quite heavily and can help explain those variations.

Housing Price Models

Introduction of Parametric vs Non Parametric

Upon exploring the variables. Its time to build a model. Our variables in total consist of: MLS area, home price, sqft, bedroom, full bathroom, three quart bath, half bath, quarter bath, year built, pool, garage space, longitude and latitude. Useful transformations observed from EDA would be for home prices and sqft. EDA revealed that the ones with the most predicting power are in: MRS Area, sqft, bedroom, full bathroom, year built, Longitude and Latitude.

For the model portion there will be 2 categories: **parametric** and **Non-parametric approach**

1. **Parametric Models**

- These are models that are built under a certain set of conditions and assumptions and only work under these set of conditions and assumptions.
- Because of their parametric nature, they offer easy interpretation.
- The most famous is linear regression
- For the parametric models: different variations of linear regression are performed and assumptions are checked against it
- In total, there are 13 linear models done and compared for the parametric results.

Parametric Model

1. Full model
2. Reduced model
3. Transformed Model
4. Reduced model with all interaction
5. Transformed model with interaction
6. M5 optimized to best adjRsquare
7. M5 optimized to best BIC
8. M5 optimized to Mallows Cp
9. M5 optimized by stepwise forward
10. M5 optimized by stepwise backward
11. M5 optimized by stepwise both
12. Lasso
13. Ridge Regression
14. Elastic Net

Non-Parametric

- Non Parametric models are model that do not need the data to meet certain parameteres and are quite flexible with how they deal with things. The problem with many of these is that as complexity of your model goes up you also lose interpretability.
- The non parametric models used are:

Non-Parametric

1. Decision Tree
2. Random Forest
3. Neural Network
4. K-nn (only cordinates)
5. K-nn (all variables)
6. K-nn (important)

The main models that will be compared are the following:

Total

1. Full model
2. Reduced model
3. Transformed Model
4. Reduced model with all interaction
5. Transformed model with interaction
6. M5 optimized to best adjRsquare
7. M5 optimized to best BIC
8. M5 optimized to Mallows Cp
9. M5 optimized by stepwise forward
10. M5 optimized by stepwise backward
11. M5 optimized by stepwise both
12. Lasso
13. Ridge Regression
14. Decision Tree
15. Random Forest
16. Neural Network
17. K-nn (only cordinates)
18. K-nn (all variables)
19. K-nn (Important Variables)
20. Elastic Net

Criteria

The different critieria looked at are:

Adjusted R-square (train), Adjusted R-square (test), MSE, MAPE for both training and test data and model coefficient number.

Linear Regression Models

LM Model 1 and 2 Full and Reduced Regression and transformation

The first model I would like to try would be a a full model encompassing all the variables. This is a basic linear model of area, price, sqft, bedroom, full bathroom, year built, pool, garage space, latitude and longitude. NOTE: Three quarter bath, half bath, and quarter bath were dropped because they added little to no value and there were consisted of less than 8% occurances.

The original model consists of all those variables area, sqft, bedroom, full bath, year built, pool, garage space, and longitude and latitude (M1). Because of EDA and analysis, pool and garage space had little explanatory power and were dropped (M2).

I checked the assumptions of linear regression on the smaller model of M2. This is because if the model with smaller variables does not satisfy the Gauss-Markov conditions, adding more variables will only make it harder to satisfy the assumptions.

The assumptions are the following:

1. **Linear Relationship**

The base assumption of running a *linear* regression is that there is a linear relationship between the explanatory variables and your data. From EDA you see a linear relationship between price and sqft, however for other variables it does not stand strongly.

2. **Your distributions to come from multivariate normal**

We see that most of our distributions do not come from multivariate normal

3. **No or little multicolinearity**.

- checking the VIF all the numbers are below 2.3 (Satisfied)

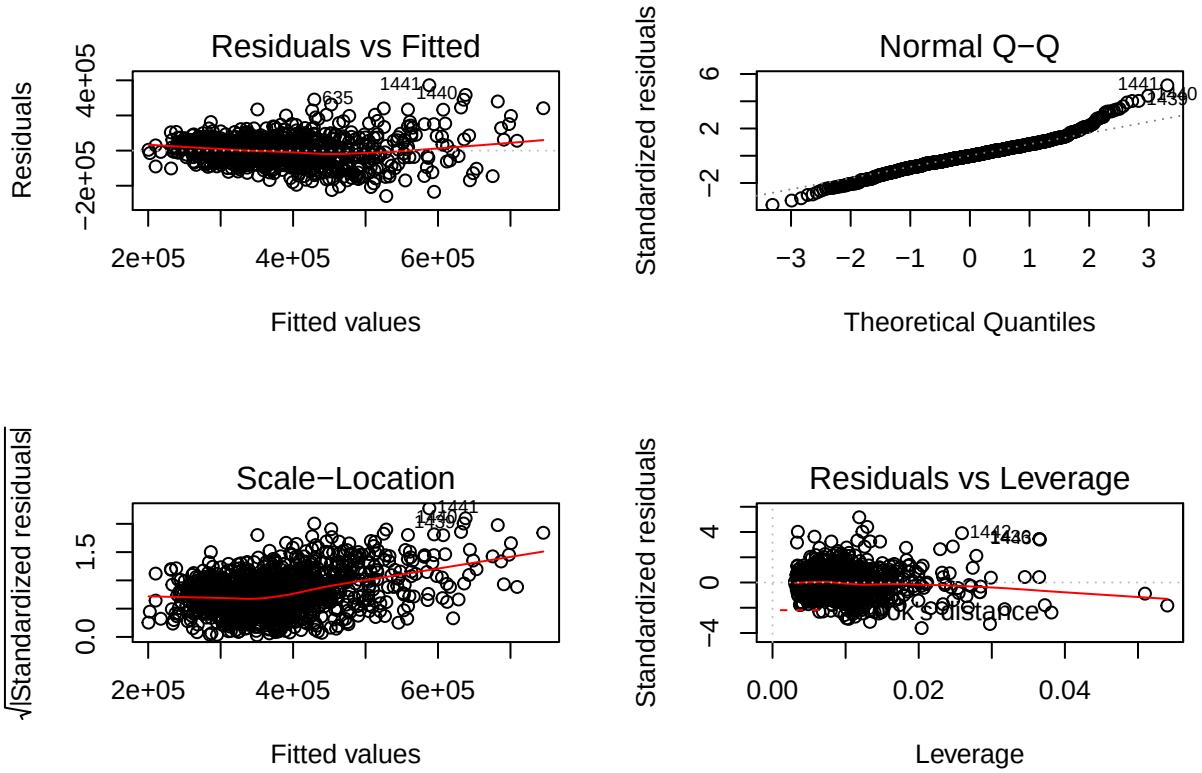
4. **Homoscedasity**

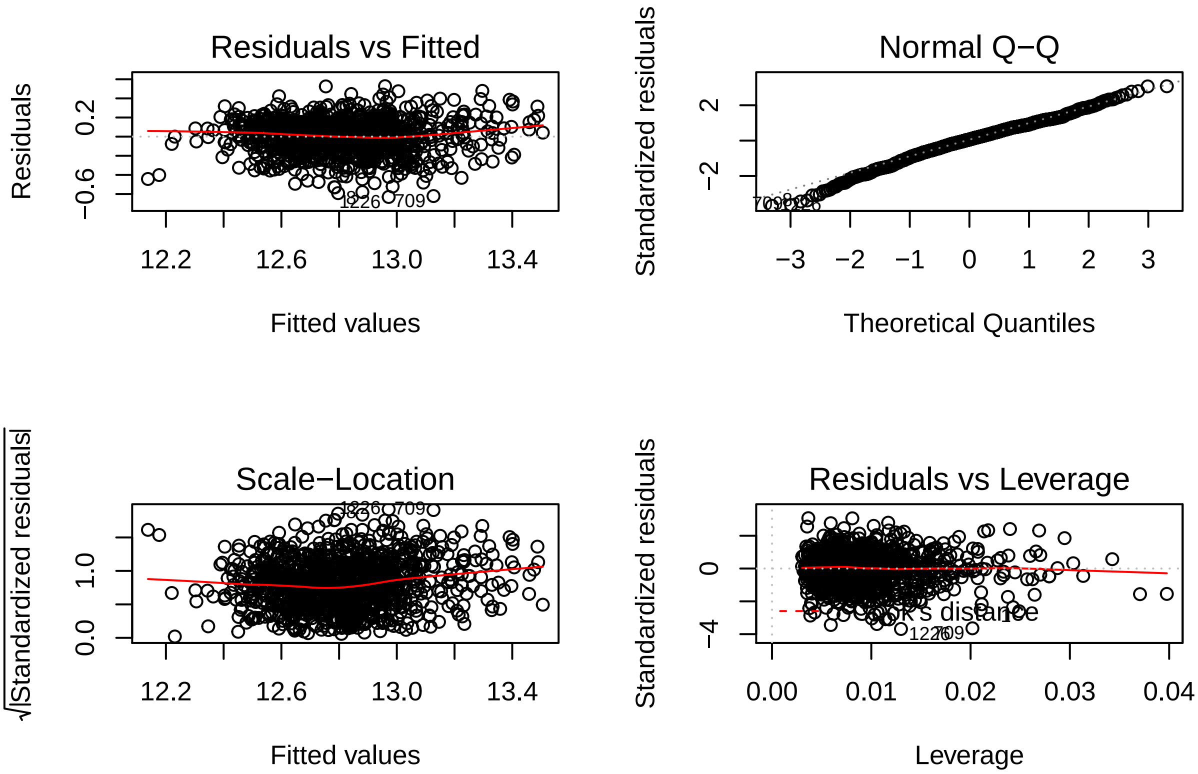
This model is heteroscedastic as the variance changes over time.

5. **Normality of Errors**

- From the QQ plot, you can see that the residuals are not normally distributed (due to fat tails)

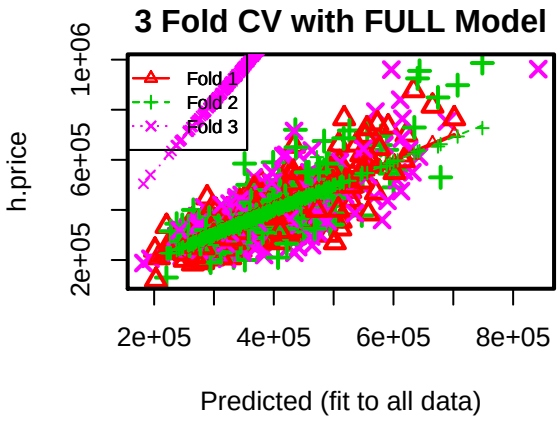
The repeat is applied to the transformed variables below and one can that althought the transformed is close it still does not follow the assumptions.

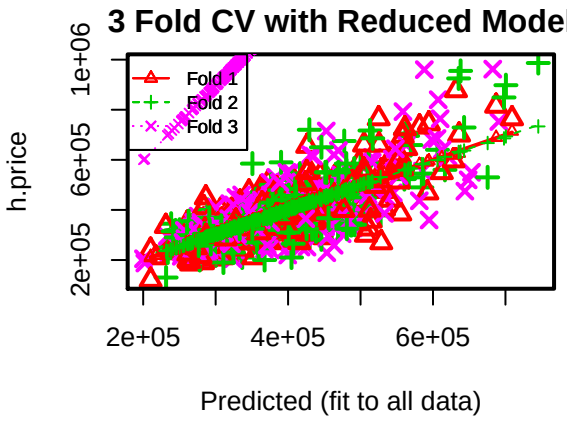


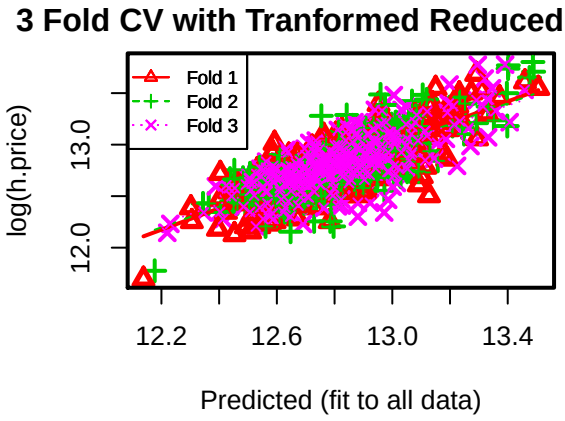



```
##
## M1/M2/M3 Results
## ================================================================================================
##                                                          Dependent variable:
##                                       -----------------------------------------------------------
##                                                    h.price                                log(h.pric
##                                             (1)                     (2)                     (3)
## -----------------------------------------------------------------------------------------------
## MLS.areaMetropolitan South          -16,696.690***          -18,921.770***            -0.051**
##                                      (5,561.066)             (5,631.607)              (0.013)
##
## MLS.areaMetropolitan Southwest        7,283.984               7,670.419                0.024
##                                      (6,248.125)             (6,361.080)              (0.015)
##
## MLS.areaMid Los Angeles             106,513.700***          105,621.300***            0.211***
##                                      (7,462.613)             (7,578.209)              (0.018)
##
## sqft                                  107.885***              118.587***
##                                        (7.939)                 (7.881)
##
## bedroom                               -1,076.085              -2,265.830               -0.010
##                                      (3,926.337)             (3,986.564)              (0.010)
##
## full.bath                           19,722.320***           18,026.460***             0.041***
##                                      (4,690.310)             (4,767.168)              (0.011)
##
## yr.built                             -321.634***              -194.912*               -0.0003
##                                       (109.704)               (108.559)               (0.0003)
##
```

```
## poolY                          154,310.500***
##                                 (50,665.780)
##
## closed.Garage.Spaces            14,763.340***
##                                 (2,679.269)
##
## Longitude                      -771,350.700***          -829,659.900***          -2.072**
##                                 (82,926.340)             (83,537.350)             (0.199)
##
## Latitude                        320,230.700***           290,745.400***           0.620***
##                                 (57,046.110)             (57,854.600)             (0.138)
##
## log(sqft)                                                                        0.422***
##                                                                                  (0.025)
##
## Constant                     -101,313,070.000***      -107,441,152.000***      -255.799*
##                               (9,449,950.000)          (9,528,829.000)          (22.689)
##
## ------------------------------------------------------------------------------------------
## Observations                          1,075                    1,075                 1,075
## R2                                    0.600                    0.584                 0.589
## Adjusted R2                           0.596                    0.581                 0.585
## Residual Std. Error          71,119.050 (df = 1063)   72,409.720 (df = 1065)    0.172 (df =
## F Statistic                144.812*** (df = 11; 1063) 166.344*** (df = 9; 1065) 169.455*** (df =
## ==========================================================================================
## Note:                                                          *p<0.1; **p<0.05; *
```







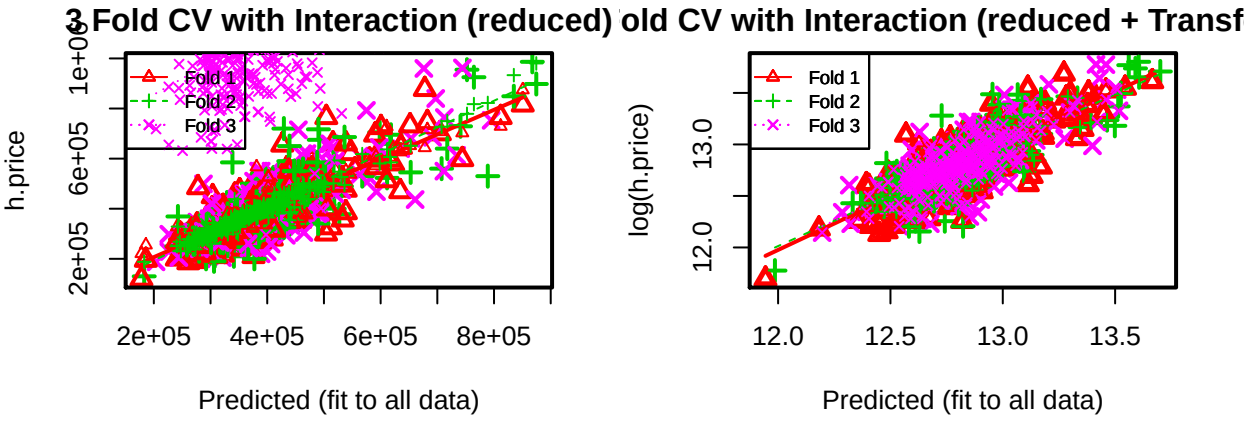
Each model was trained and tested on a 3 fold cross validation and the results are plotted for better visualization.

Checking the parameters of a linear regression, all models have been shown to fail the assumptions and cannot be applied with high confidence. Even transformed it does not comply. For the sake of the report, well have to assume that model would be able to stand on its own; even though it is incorrectly trained and does not satisfy the parameters needed. For that reason, additional models are created to compare and tested.

Model 4,5 Interactions

Model 4 and 5 are models that include the interactions of all the different types of variables from model 2 and 3. This is to help capture any sort of non-linear interaction effect going on. This was done after conducting a reset test and determining that there is some a non-linear relationship with the variables.

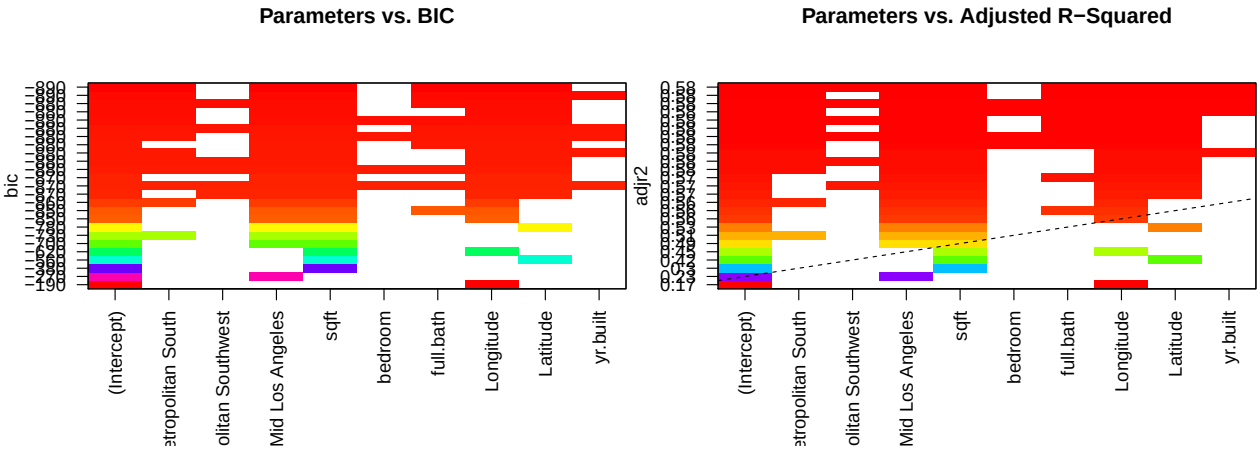
After running a full scale interaction model, the relationships between sqft and area came out to be statistically signficiant as hypothesized. Although this part of the model came out correct. It is quite naiive to assume that all these interactions are needed. Therefore, the best approach would be to try dropping them. The way that I would like to drop them is by using stepwise regression and regularization.



Before moving on to stepwise and regularization. I wanted to try and optimize the single model by regards to the three paratmeters (Adjusted R square, BIC, and Mallows CP)

R-square and friends (Model 6-8)

```
## Reordering variables and trying again:
```



```
##
## adj R-square and Friends Models
## ================================================================================================
##                                                  Dependent variable:
##                                     ---------------------------------------------
##                                                        h.price
##                                          (1)                (2)                (3)
## ------------------------------------------------------------------------------------------------
## MLS.areaMetropolitan South          -24,180.670***     -20,297.680***     -24,180.670***
##                                       (5,632.540)        (5,589.111)        (5,632.540)
##
## MLS.areaMetropolitan Southwest        5,227.268          7,753.431          5,227.268
##                                       (6,467.834)        (6,366.027)        (6,467.834)
##
## MLS.areaMid Los Angeles             113,498.100***     105,945.400***     113,498.100***
##                                       (7,466.733)        (7,582.444)        (7,466.733)
##
## Longitude                           -969,241.400***    -839,124.500***    -969,241.400***
##                                       (80,911.660)       (83,394.170)       (80,911.660)
##
## Latitude                                               313,232.200***
##                                                         (56,651.330)
##
## sqft                                  132.119***         115.803***         132.119***
##                                         (5.535)            (6.923)            (5.535)
##
## full.bath                                               15,787.980***
##                                                          (4,463.743)
##
## Constant                          -114,443,318.000***  -109,700,785.000***  -114,443,318.000***
##                                     (9,570,279.000)    (9,461,390.000)      (9,570,279.000)
##
## ------------------------------------------------------------------------------------------------
## Observations                            1,075              1,075              1,075
## R2                                      0.566              0.583              0.566
## Adjusted R2                             0.564              0.580              0.564
## Residual Std. Error              73,810.460 (df = 1069)  72,468.130 (df = 1067)  73,810.460 (df =
## F Statistic                     279.355*** (df = 5; 1069) 212.995*** (df = 7; 1067) 279.355*** (df =
## ================================================================================================
## Note:                                                          *p<0.1; **p<0.05; ***
```

With adj R square, BIC, and Mallows CP three different models were represented. With adjusted R square

Since we see a signficant drop without the interaction terms and by indications from a reset test. It would be best to keep the interaction terms. However to help figure out which interactions to keep we look to stepwise and penalizing regressions.

Stepwise Regression (9-11)

With Stepwise Regression, three different methods were constructed. One backward, one forward, and one using both method. In statistics, stepwise regression is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure. In each step, a variable is considered for addition to or subtraction from the set of explanatory variables based on some prespecified criterion.

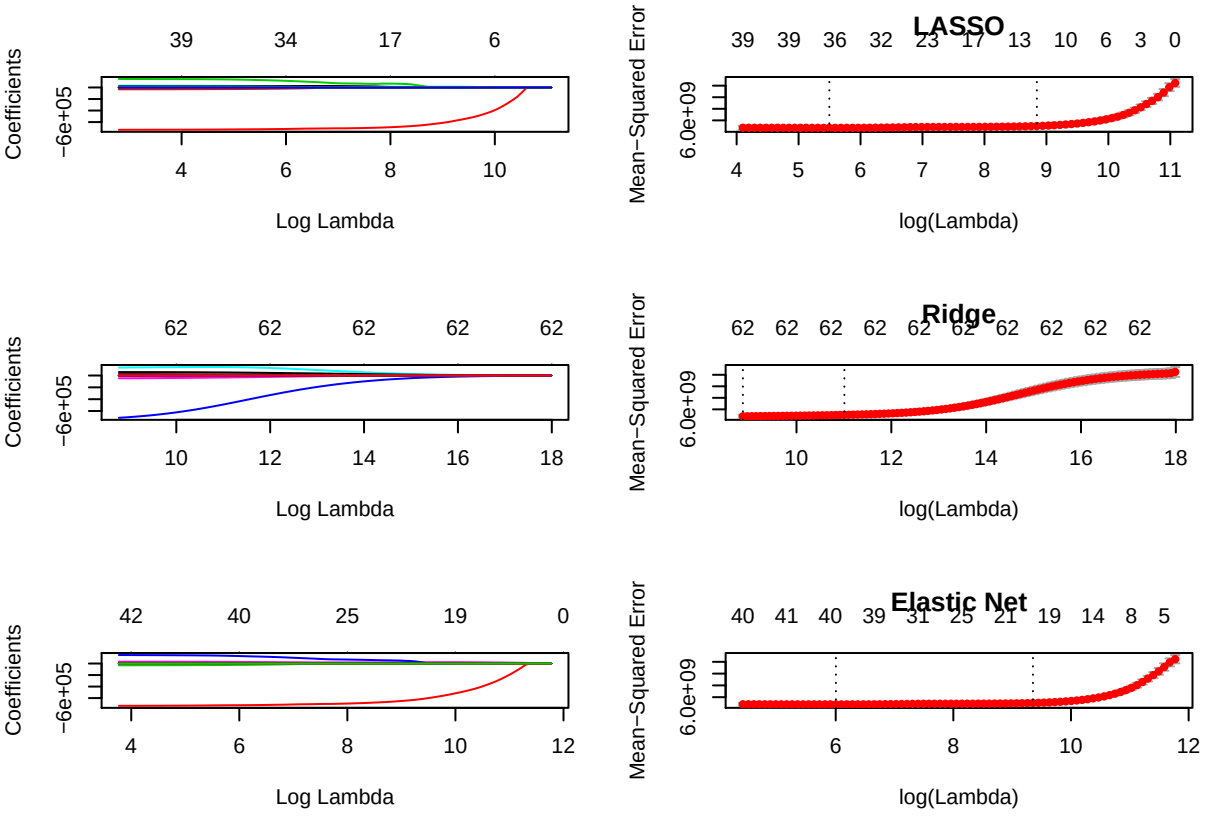
Stepwise forward you start out with nothing and keep tacking on until you cannot improve R square any

more. Stepwise backward you start out with you full model and remove variables. While Stepwise both, uses both.

The problem with stepwise is that it does not coerce any of the useless variables to zero. As a result, many of the terms still came out to be highly significant. Therefore regularization might be better.

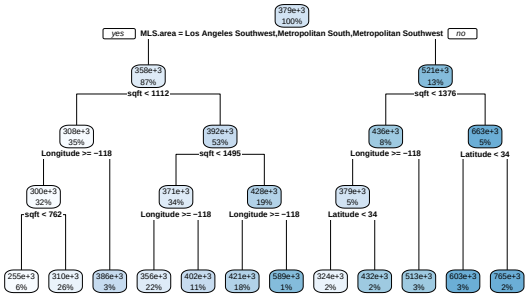
Regularization

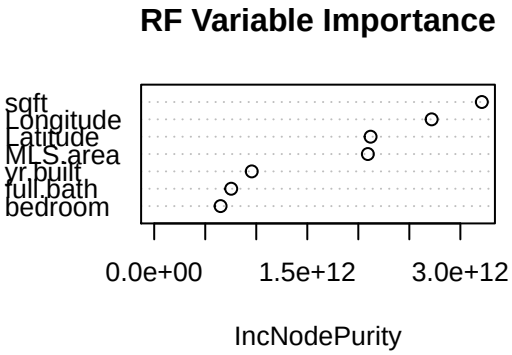
Regularization coerces the values to zero or near zero by a penalty function. I have ran three different regularization techniques. The first one is lasso, then ridge, then elastic net. Lasso coerces variables to zero therefore it was easy to print out the meaningful results. It says that there are 11 variables to consider. Elastic Net was able to break it down to 17. While Ridge as expected couldn't reduce it down. Printed below are lasso's coefficients. So far this is the most promising model.



feature	coeficient
(Intercept)	-8.296852e+07
full.bath	1.985936e+03
Longitude	-6.303338e+05
MLS.areaMetropolitan South:sqft	-1.027657e+01
MLS.areaMid Los Angeles:sqft	6.063654e+01
MLS.areaMid Los Angeles:full.bath	1.314302e+04
sqft:closed.Garage.Spaces	7.220988e+00
sqft:Latitude	2.491370e+00
full.bath:Longitude	-4.071795e+00
full.bath:Latitude	1.138639e+02
Longitude:Latitude	-2.152017e+03

Decision Tree





A highly used regression in machine learning is a regression tree. These are splits within the data than can help classify it. Although a single regression tree is inadequate at properly modeling. It is a great explanatory power and provides a pleasant visualization. To hep boost its predictive power, I did an ensemble method called Random Forest combining it with 500 other trees each one with its own splits. The random forest has a variable importance plot shown to the right. Which shows the importance of each variable. The main ones are sqft, area, and Longitude and Latitude. The single tree also emphasizing that these are the top variables. The results agree with what the other models were doing.

Neural Network

The last machine learning technique that I will try to solve the problem with is approaching it using a neural net. Each node gets a linear combination of the previous number of nodes and iterates through the process. In this case because the focus is on regression a linear activation function was chosen.

```
## [1] -1.055494562
```

Knn

The benchmark to the other models will be K-nn. This is essentially to mimic real world price valuation method such as comparable. To construct this: I approached k-nn through 3 criteria. One k-nn model simply only uses GPS cordinates. The results of the that model and the corresponding k are below as the first one. The second one is k-nn using all the variables that we had. The results are displayed in the second table. The third table are using the important variables of area, longitude, sqft, and latitude.

k	RMSE	Rsquared	RMSESD	RsquaredSD
5	97101.83815	0.2771900241	9771.08514	0.0708353027
7	95759.69772	0.2873792274	10456.95910	0.0838792646
9	94483.29153	0.2982226709	10929.54935	0.0940732295
11	94033.26967	0.2996141099	11961.67001	0.0996104986
13	93461.92475	0.3070189598	11995.23835	0.1033181276
15	93523.44373	0.3044117338	11662.19058	0.0925484491
17	93002.65376	0.3111190325	12041.56070	0.0973735907
19	92350.83313	0.3186826073	11570.78568	0.0894859651
21	92042.96326	0.3231676543	11624.78504	0.0841128639
23	92258.11276	0.3196976972	11431.48973	0.0837068886

k	RMSE	Rsquared	RMSESD	RsquaredSD
5	71186.42660	0.5946216816	5861.833988	0.0368494371
7	70699.90387	0.5989050666	6434.774103	0.0358078640
9	70276.97194	0.6031529725	5806.354292	0.0433588530
11	70321.81771	0.6043208547	5555.736768	0.0403958849
13	70224.99904	0.6073065847	5246.652479	0.0345482461
15	71032.77338	0.5992540221	5348.279845	0.0291641110
17	71690.05774	0.5929956822	5549.460741	0.0225792664
19	71844.14101	0.5919849434	5432.189140	0.0162474409
21	72141.27894	0.5900283075	5289.882605	0.0166048424
23	72534.07423	0.5858858082	5506.801048	0.0145678108

k	RMSE	Rsquared	RMSESD	RsquaredSD
5	71628.09562	0.5958625151	5780.891684	0.0247498395
7	70446.08577	0.6086376640	5233.384140	0.0270782472
9	69116.33265	0.6226945006	5751.657387	0.0209607515
11	69595.31653	0.6176428283	6182.358800	0.0188065878
13	70268.28281	0.6105304232	6388.240519	0.0312671248
15	70765.68677	0.6048990830	6833.763904	0.0417523619
17	71082.00980	0.6018171778	6562.178270	0.0399206162
19	71555.95279	0.5977069909	7050.329322	0.0397708898
21	72083.73707	0.5924163836	6887.452593	0.0357958496
23	72299.14664	0.5909243273	6759.708546	0.0390574452

	Adj_R2	Train_RMSE	Train_MAPE	Test_RMSE	Test_MAPE	Coefficient #	Adj_I
Full_Model	0.5956222693	70720.99619	0.1413354737	74970.96047	0.1424631027	12	0.575
Red_Model	0.5808117832	72072.14688	0.1450879574	73969.95452	0.1418539857	10	0.589
Transformed	0.5853429855	69591.76806	0.1361106637	71768.12599	0.1338829402	10	0.613
Red_Interaction	0.6561214194	64258.49257	0.1272796695	68414.10771	0.1303654021	43	0.597
Trans_Interaction	0.6241916174	64569.79601	0.1246954597	68954.94105	0.1291381586	43	0.591
AdjR2_optimized	0.5644368529	73604.19006	0.1471478933	75654.80263	0.1473142821	6	0.577
BIC_optimized	0.5801352688	72197.97709	0.1452923880	74054.40267	0.1416427709	8	0.591
MallowsCP_opt	0.5644368529	73604.19006	0.1471478933	75654.80263	0.1473142821	6	0.577
Step_Forward	0.6559442342	64554.70590	0.1274275926	70388.93456	0.1317608386	34	0.590
Step_Back	0.6561214194	64258.49257	0.1272796695	68414.10771	0.1303654021	43	0.597
Step_Both	0.5799373687	72181.14194	0.1455275147	74062.99771	0.1419638968	9	0.590
Lasso	0.5878499085	69663.62644	0.1295627021	74015.12518	0.1379445994	63	0.483
Ridge	0.6035213764	70059.78891	0.1320581089	72379.41385	0.1386051758	11	0.605
Decision Tree	0.6246830552	68388.32132	0.1318933214	83703.64371	0.1606340211	4	0.486
Random Forest	0.9138115983	32726.39455	0.0633566048	66637.50681	0.1251674184	7	0.670
Neural Network	0.6532867091	65577.06412	0.1247684603	168400.86279	0.3893351009	9	
K-nn (cordinates)	0.3863966317	87565.72144	0.1604868420	91515.24934	0.1735306990	0	0.395
K-nn (all)	0.6554237161	65619.50563	0.1229983020	71331.90195	0.1337116588	0	0.631
K-nn (important)	0.6991893013	61310.78855	0.1163842194	69089.98357	0.1331188340	0	0.654
Elastic_net	0.6039770495	69821.81490	0.1317254496	72440.46666	0.1381693692	17	0.595

Conclusion

```
##
## Winning Basic Results
## ================================================================================
##                                          Dependent variable:
##                                 ---------------------------------------------
##                                       log(h.price)                h.price
##                                           (1)                       (2)
## --------------------------------------------------------------------------------
## MLS.areaMetropolitan South              -0.051***               -20,297.682***
##                                          (0.013)                  (5,589.111)
##
## MLS.areaMetropolitan Southwest            0.024                    7,753.431
##                                          (0.015)                  (6,366.027)
##
## MLS.areaMid Los Angeles                 0.211***                 105,945.381***
##                                          (0.018)                  (7,582.444)
##
## bedroom                                  -0.010
##                                          (0.010)
##
## sqft                                                               115.803***
##                                                                     (6.923)
##
## full.bath                                0.041***                 15,787.982***
##                                          (0.011)                  (4,463.743)
##
## yr.built                                 -0.0003
##                                         (0.0003)
##
## Longitude                                -2.072***                -839,124.518***
##                                          (0.199)                  (83,394.174)
##
## Latitude                                 0.620***                 313,232.167***
##                                          (0.138)                  (56,651.327)
##
## log(sqft)                                0.422***
##                                          (0.025)
##
## Constant                                -255.799***             -109,700,784.600***
##                                          (22.689)                 (9,461,390.108)
##
## --------------------------------------------------------------------------------
## Observations                               1,075                     1,075
## R2                                         0.589                     0.583
## Adjusted R2                                0.585                     0.580
## Residual Std. Error                  0.172 (df = 1065)      72,468.129 (df = 1067)
## F Statistic                       169.455*** (df = 9; 1065) 212.995*** (df = 7; 1067)
## ================================================================================
## Note:                                                *p<0.1; **p<0.05; ***p<0.01
```

In the above is the table for the results comparitively.

Compared to our basic models of full model, reduced model, and Transformed. Transformed takes the best

cake with reduced variables and slightly better.

Comparing the non interaction models optimized by Rsqaure, BIC, and Mallows CP. The best one was the one that got optimized by BIC. Two two basic regressions are posted below. Of these two, the transformed model did best.

If you factor in interaction terms through stepwise and regularization, there is about a less than 1 percent reduction mean absolute percentage error. Some basic observations are: The decision tree was overfitted thus getting 97 percent r square on training and about 48 percent on testing data. Either way it wasn't that off. What's more is that by combining many of them and getting a random forest you are left with the best model. In order the best models were 1) Random forest 2)transformed reduced interaction (all) 3) transformed interaction (all) 4) and transformed.

How does it compare to benchmark?

K-nn was set aside as the benchmark model. The one including only the coordinate was worse than all of my machine learning models (except neural network... don't know what happened over there). Meaning that modeling did have some power over it, however the bigger benchmarks are the knn with all the variables and the most important ones. Out of those benchmark everything failed but random forest. To test whether or not the difference is signficant, a diebald mariano clark west test was conducted.

```
## Warning: package 'forecast' was built under R version 3.3.2
```

```
##
##  Diebold-Mariano Test
##
## data:  m18_predict_test - R10.test$h.pricem15_predict_test - R10.test$h.price
## DM = 1.5406293, Forecast horizon = 1, Loss function power = 2,
## p-value = 0.06229346
## alternative hypothesis: greater
```

```
##
##  Diebold-Mariano Test
##
## data:  m19_predict_test - R10.test$h.pricem15_predict_test - R10.test$h.price
## DM = 1.1921475, Forecast horizon = 1, Loss function power = 2,
## p-value = 0.117129
## alternative hypothesis: greater
```

Unforunately, both models indicate that the Random Forest is not statistically different in predicting power than a K-nn using the best predictors.

Lastly, from everything that was run it can be said with utmost confidence that the primary features of the house contribute to the price are sqft and LOCATION, LOCATION, LOCATION!!!!!

Improvements and Limitations

As a researcher, I acknowledge my own limitations and the limitations of this research. There are many ways to improve on this study and possible find a better model than comparables in real estate.

Improvements

- Run everything transformed

As shown transforming the variables to meet the parameters of linear regression did a lot to transform perfomance. Perhaps doing a combination of regularization and transformation can help. Or applying transformed data to the different machine learning models.

- Improved Models

There are a few models that were not tested and have had remarkably more success in this area such as xgboost and hedonic regression. Furthermore, a possible ensemble of different models here could possibly beat the real estate forecast.

- Economic variables

Another way to improve the study is to add economic and demographic variables of the neighborhood the house is as inputs for the model. That would greatly increase the explanatory variable.

- Tackling Rent prices

This was quite fun and I would love to apply the same methodology to try to predict rent prices as that is more of tricky than trying to properly evaluate the price of your home.

- Sentiment Analysis (Google Trends & Twitter)

Something I wanted to try to implement was examining the impact of google trends and sentiment analysis on twitter with the associated real estate area and seeing the impact that would have.

Limitations

As you saw from the summary, the greatest price of the home was under 1 million dollars and the model loses stability with higher ended numbers. The models used here are limited only to the Los Angeles area with the provided data points and cannot handle outliers well. Furthermore, the data will only work for the selected year and does not hold any time component to it. Ideally, I would love to revist this project and implement a fixed effect and random effects component to it.

References

- StackOverflow
- https://beckmw.wordpress.com/2013/11/14/visualizing-neural-networks-in-r-update/
- https://drsimonj.svbtle.com/ridge-regression-with-glmnet
- https://www4.stat.ncsu.edu/~post/josh/LASSO_Ridge_Elastic_Net_-_Examples.html#generate-data-1
- https://web.stanford.edu/class/cs221/2017/restricted/p-final/ianjones/final.pdf
- https://www.kaggle.com/c/house-prices-advanced-regression-techniques
- https://towardsdatascience.com/create-a-model-to-predict-house-prices-using-python-d34fe8fad88f
- https://nycdatascience.com/blog/student-works/predicting-house-prices-using-machine-learning-algorithms/
- https://www.r-bloggers.com/a-data-scientists-guide-to-predicting-housing-prices-in-russia/
- http://rl.cs.mcgill.ca/comp598/fall2014/comp598_submission_99.pdf -http://people.duke.edu/~rnau/testing.htm