Beach Market Behavior

- A Data Driven Real Estate Study -

By: Dan Fugardi

January 6th, 2020

USC Marshall

School of Business

Table of Contents

I. **Introduction:**

In this report, we will be exploring an analysis of the real estate markets: Redondo Beach, Manhattan Beach, and Hermosa Beach, located in Southern California. These three beach cities are iconic areas between Los Angeles and Orange County California, and while desirable areas, they are considered relatively expensive according to the rest of the country. They are markets in which one could benefit greatly from understanding the full statistical behavior of prices. The nature of a Southern California beach town can be whimsical, therefore how the common person might make buying or selling decisions can be tricky to understand season-to-season, if the basis for a house's price is not fully understood.

The parameters provided were as follows: Home Type, City, Beds, Baths, Size (SF), Lot Size, Year Built, Parking, Days On Market, and of course, Price: in this case, List Price.

The primary objective of analyzing this data is knowing the proper price at which a seller or a listing agent might want to list a house in order for it to sell in an acceptable amount of time. The Days On Market timer stops once the property is put into escrow. A real estate sale target is typically considered great if the sale happens under a month. This is generally known among experienced agents as the most important window in the real estate sales process.

Secondarily, this data, if able to be used as an accurate indicator for price, could also be used to help buyers understand what the market should be for the purchase of a particular house. While this would apply slightly more to the investor/developer type, it will also apply to the owner/user.

That said, people buying a home to live in are likely to be less concerned about market price sensitivity than an investor/developer, whose sole purpose and necessity is based only around making the correct financial decisions. An owner/user is more likely to pay top dollar for a house that is inline with the vision they had dreamed about.

The emotional component in residential real estate can account for far more variation in prices, especially when these emotions are in place of professional knowledge. This is a big factor that differentiates real estate business from industries that have little to no emotional purchasing component, such as an industrial B2B fabric purchaser knowing the exact market standard for a square foot of mid-quality silk.

The importance of this information is not to suggest the obvious correlation between emotions and price anomalies, but the significance of understanding what *drives* this behavior is necessary because the result is the possibility of measuring and predicting the skewed data as well. The net result is certainly higher difficulty in measuring skewed data. This "higher difficulty" is statistically represented as a lower R^2 ("R Squared"). The R^2 value is represented as a decimal, in place of representing a percentage of the data that is supporting the trend, which gives us our predictions. For example: An R^2 of .75, tells us that 75% percent of the data is supporting the linear slope across the X & Y Axis, such as Price & Size. This also says that if we price something according to that trend, we have a 75% of being accurate.

Regarding the more emotionally-driven data, assumptions will include that there will likely be a correlation made between the age / quality of materials used in the house, and people being motivated to make purchases outside of what market data might otherwise dictate.

A secondary or even tertiary variable that is missing and could be very important to incorporate at a later time, and one that could also be easily overlooked in full analysis, is the correlation between how far UNDER priced a house may be, and the ultimate sale price being pushed in the upward direction. There is a [not entirely unexplainable] phenomena in real estate, as in any auction-like environment, which is that of a higher sale price taking place because of an under-priced asset. As the price of a house in this case is clearly a "deal", the traffic of those viewing the house dramatically increases and often leads to a bidding war, which can lead to inflated prices. Today's particular study is focused on List Price, not Sale Price, and is attempting to make a determination of a time-based result, not a price-based result. That said, in a supplementary study to further round out the big picture, a focus on the effects of Sale Price would be something an analyst would also want to consider.

More importantly, and more relevant to this study, the data provided also does not provide information based on internal details of the house such as type of surfaces used, cabinets, flooring, windows, hardware, and in this case, pools are also not included. My experience in real estate dictates that the aforementioned variables are hugely important. While my goal is to find out if a correlation can be made from market stats only, I have, from the onset of this exploration, had the suspicion that the lack of material data could heavily impact this study.

II. <u>Hypothesis & Proposed Objective:</u>

My hypothesis was that a multiple regression analysis could be run on all variables *except* Days On Market, in order to find predicted values for what a house should have been priced, without any consideration for it's Days On Market before selling. In concept, we would start with previous List Price used, which was supported by information the agents had access to at the time, in order to estimate what they thought the house price should have been. As real estate professionals aren't typically data experts, we can not make the assumption that their basis for pricing is near 100% accurate.

That said, these agents are industry professionals and do at the least typically know their micro-markets relatively well. While the information for forming these List Price assumptions are not the same as the results of a regression model prediction, the basis is an aggregate of price-per-square-foot measurements of houses in the surrounding area, and those comparables are often relatively accurate. While one can sport what "the data says" would be considered absolutely correct, the *true* market price for a house is: one in which the market accepts the price, or in other words, the price in which the house sells, and does so in a short period of time.

Based on that universal concept, the longer it takes for a house to sell, the greater of a measure that the house was priced “incorrectly.”

Subsequently, my first target focus was such that we start by finding the reasoning for the List Prices used based on the other components of the house. Accordingly, an assumption is that we will be left with what the data tells us should be the true price for the house, and how far away from these theoretically true data-backed prices the originally used List Prices were.

My objective is then to determine if enough information exists within this strictly online market data set, in order to be able to make accurate determinations on (1) what dollar amount a house should be priced at, (2) what the delta is between the price the agent estimated and the price “correct” data-backed number at which the house should have been priced, and lastly, (3) if this delta showing how far the target price was missed, correlates to the amount of time that is takes for the market to absorb the house.

III. Assumptions & Observations:

One of the initial observations that was made in the first model that included all variables, was that Days On Market was one of only three variables that was most heavily weighted. House Size and Year Built were the first two, which is highly logical.

Days On Market was weighted heavily as well and, as explained, this would normally make sense. The Days On Market would typically be correlated with price, but there is no data showing Reduced Prices or Sale Prices, and therefore, the goal is to try and find a way that the List Price itself, inclusive of the remaining variables available, will let us draw a conclusion about the accuracy of that List Price, and subsequently, a correlation to Days On Market.

Before going deeper into multiple regression observations, let's look at basic linear observations that might seem obvious to a real estate professional, and even common sense to the general public. It’s possible that, in this application, real estate is something everyone encounters and is therefore somewhat familiar with. What if we were explaining this to an alien, or more likely, another industry was being observed, and the assessor had no experiential understanding of the industry? Furthermore, after looking at the “obvious”, is it possible that the slope of the correlation can be heavily affected and therefore is hardly a given? Let’s explore and challenge.

Assumption: There is a definite correlation between Price and Size. In other words, a larger house will generally mean a more expensive house: True or False?

My assumption as a real estate professional, as I would assume most would answer, would be that in general the above statement is true: Houses become more expensive as they become bigger in size.

I then took a closer look at this concept:

Below we will look at graphs of how price relates to size in three perspectives: Two graphs, which are separate groupings of the three beach cities, and one graph is analysing every house together, in the data set. The two groupings are broken down by (1) Manhattan Beach & Redondo Beach combined, and (2) Hermosa Beach alone.

Before going further, it is necessary to explain the reasoning for breaking apart the groupings:

In real estate, the first thing that has to be assumed before doing a comparative analysis, is that the houses being compared, are at a minimum, in the same “market”. A market in real estate is an area in which house sales behave and trend similarly with variations in Price taking place *only* due to basic variations in the houses components, and therefore the houses in such microcosm trend in a fashion proportional to those component variabilities.

Just because there are three separately named cities, does not mean they are three different markets. This could all be considered one market, or there could be two markets among the three cities, and yes, there *could* be three separate markets. So how does one figure this out with data? If one studied each individual house sale and compared each and every house to one another until becoming familiar with the metrics, it would be a way to achieve estimates, but the power of using regression formulas is that one can simply run a multiple regression, and in the results, the “t value” will be produced for the necessary variables that will lead us to know what to look at further based on weighting. The lower the t value, the greater the significance.

It took running two subsequent models to determine that there are two markets: Redondo Beach & Manhattan Beach are one market, where Hermosa Beach is it’s own market.

This was learned by the fact that the regression outputs viewed, of all the data provided, showed that houses in Hermosa behave in a category of its own, versus houses in Manhattan & Redondo, which behave collectively, in a similar fashion. In other words, Manhattan & Redondo Beach are one market, whereas Hermosa behaves differently, and is therefore it’s own market.

“t value” is used to determine this divide. A t value indicates the significance of a test variable. While this is an abridged explanation, understand that a “significant” t value is that of a billionth.

I used the “R” Project program, and I learned from discovering when a logistical regression was run on all variables involved, Manhattan Beach & Redondo Beach were included in the output, with high t values, despite those variables not manually being singled out by myself in the input.

I learned that the formula output in R produced this to show that Manhattan and Redondo Beach had high (bad) t values for the purpose of explaining that, individually, they are not significant. In other words, it does not matter if the house is in Redondo or Manhattan, whereas it does matter if a house is in Hermosa, verses either Redondo or Manhattan. In other words, it is important to know if a house is in Hermosa or not, and can otherwise be in Redondo or Manhattan, but that will not significantly matter.

Below is how this was represented in the model.
*Observe the high t values in both models:

Model 1

CITYManhattan Beach	0.0226
CITYRedondo Beach	0.1195

Model 2

CITYManhattan Beach	0.07276
CITYRedondo Beach	0.68736

Subsequently, when I combined Redondo and Manhattan into the same group, which I named “NotHermosa” and ran the Model 2 formula again, the formula did consider the pairing and delineation of Manhattan & Redondo as it’s own market, separate from the Hermosa market, thereby giving the new paired cities a t factor of significance.

Model 2 (Revised)

NotHermosa	.00000012 ***

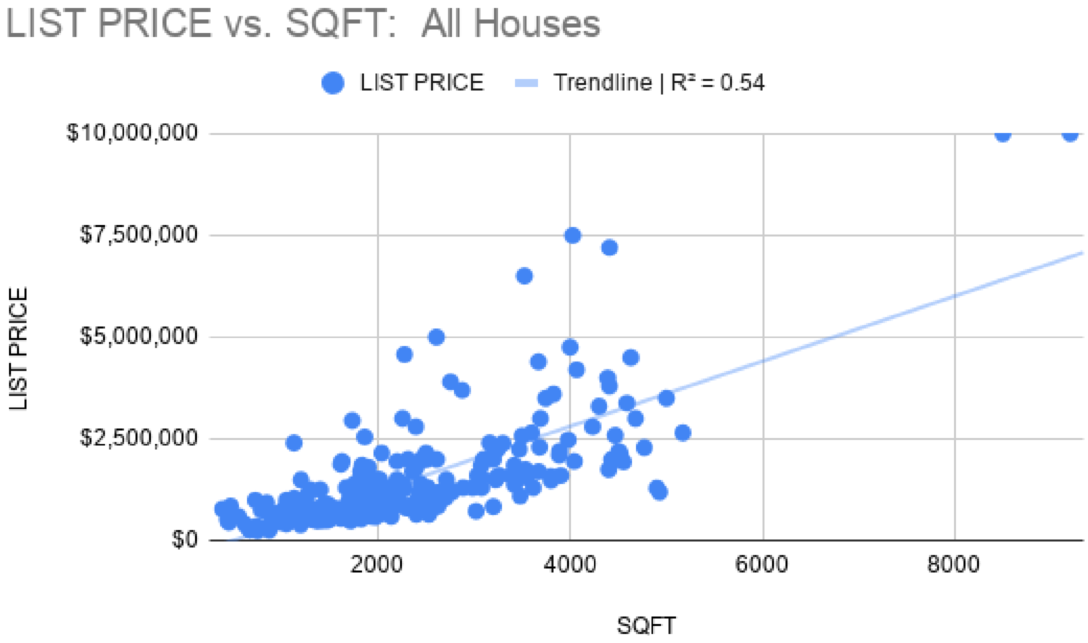
Although this is merely a very broad and very partial conclusion, what this does tell us is that upon further analysis, we need to look at these two markets separately: Manhattan Beach + Redondo Beach as one market, and Hermosa as another. The reasoning behind this is because we know that as we assess the data going forward, the more granular conclusions will be more accurate if analyzed individually as the trends and patterns will be more consistent, and will allow us to draw more accurate results, when compared to similar houses that would be affected at similar rates.

When coming back to our initial assumption, now that we know that there are two markets, the concept that we seek to find will still be looked at across both marketplaces: does Size have a linear correlation with Price? That said, with this new information, the correlation, albeit a heteroscedastic line in both scenarios, will behave slightly different when isolated to a market more relevant to the subject house. So we must break apart the data into their respective markets that we just learned about, as the data tells us to do, before moving forward.

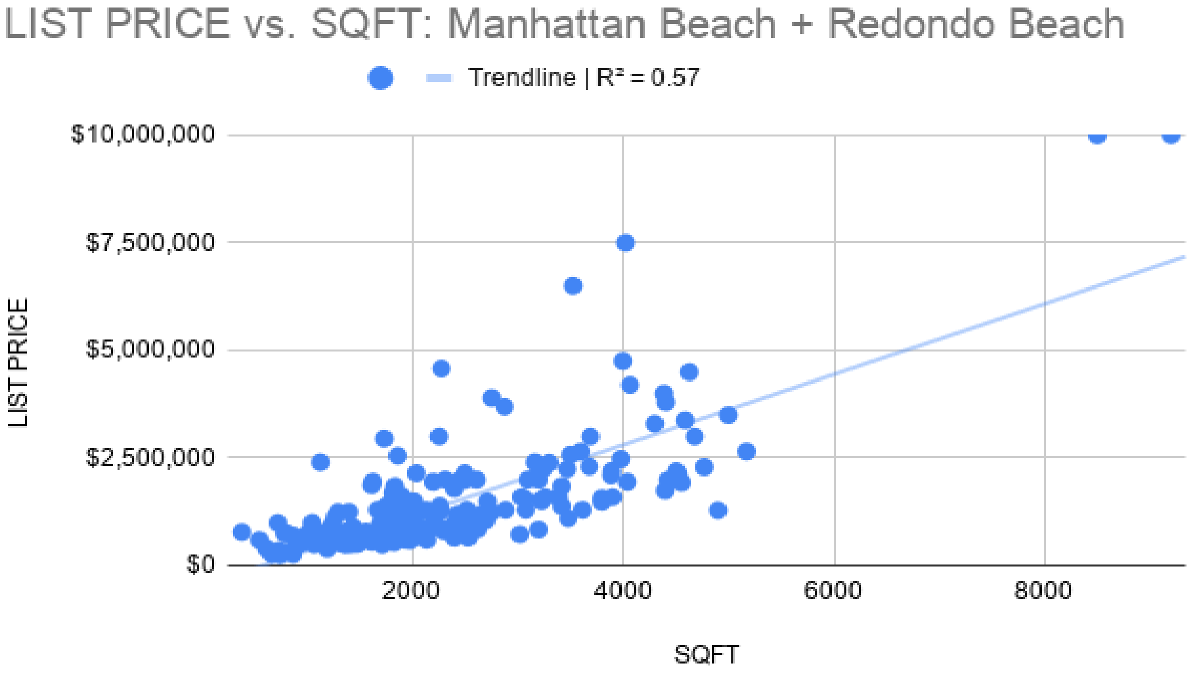
We will therefore compare Hermosa Beach against Manhattan + Redondo Beach. We will also look at the total data as a whole, so that we can see a relative comparison in behavior when taking a broader look versus separating out the analysis by market.

We’ll want to note that the R^2 figures, although not extremely strong between 50 and 60%, are all hovering around the same place and are therefore ok to make relative comparisons.

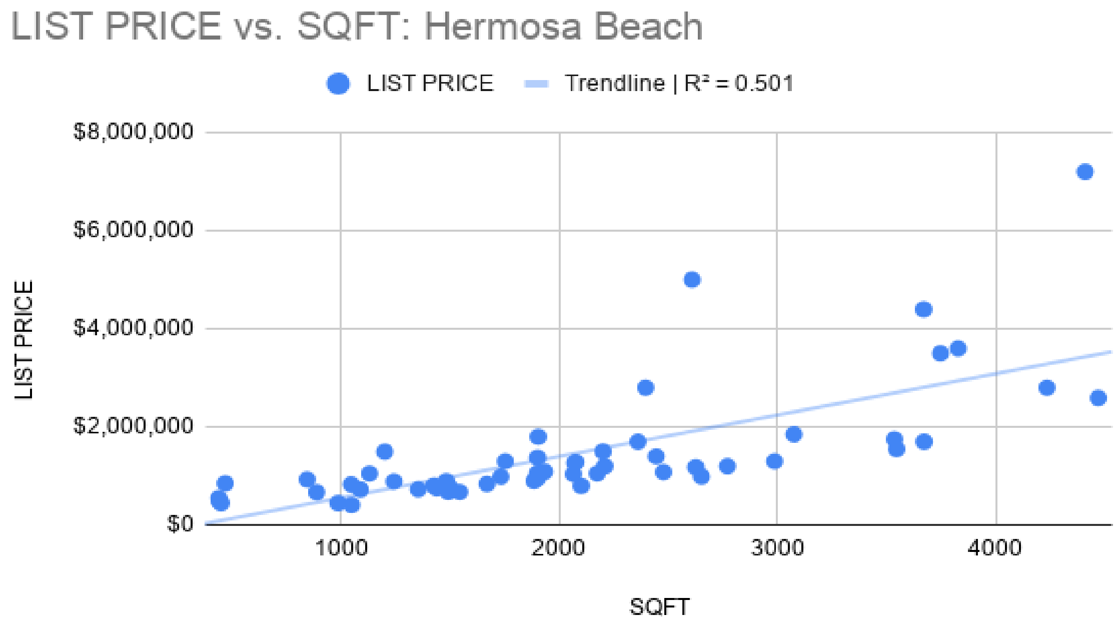
Below are All Houses: The R^2 shows that 54% of the data supports the trend, and there is a decently heteroskedastic slope giving us a general correlation between Price and Size.



The below look at the Manhattan + Redondo Beach market gives us something similar: The R^2 shows that 57% of the data supports the trend, with a similar trend line / similar correlation.



Now when we take a look at the Hermosa Beach market alone, the results are a bit different: The trendline is slightly more homoskedastic, and thus the relationship between size and price do not correlate quite as well. The R^2 is also lower at 50%, which is logical as there is more anomalous data points in this market that do not follow a predicted trend.



So what does this mean?

In beach towns, in particular, a house near or on the beach will have a hugely different price-per-square-foot value, than a house that is not near the beach. The difference between a beach house and a house inland is sometimes so big that a beach house half the size of an in-land house, could cost twice as much! In addition, considering the given that there is far less beachfront land than there is non-beach front land, this could and often does also skew price behavior due to supply and demand laws. The result of both of these concepts creates huge variability in price that is less normalized and harder to predict, and obviously also is in no way consistent with the behavior of similar in-land houses.

What the above data tells us, is that knowing the general concepts just mentioned, at a minimum, an important and much needed data point that is not included in the current data set, is distance from the beach. In a professional analysis, further discovery would be needed once the extent of our available conclusion could be found, and the further analysis would require seeking and acquiring new data that we learn we need from this study. We know now that distance from the beach will be one of those new data requirements.

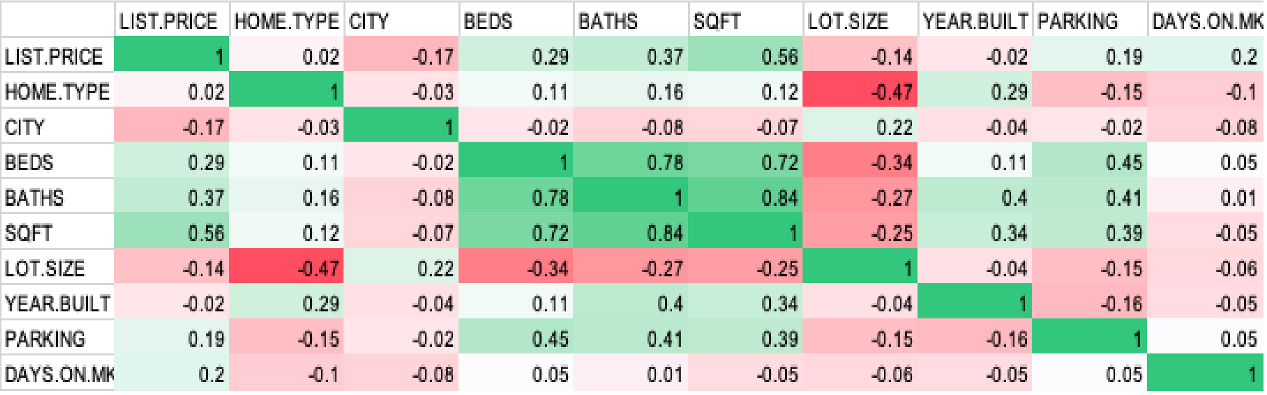
IV. **Descriptive Statistics:**

Now we will take a look at the means and standard deviations across all of the data provided. This will give us a better understanding of how the other variables affect this study.

Means and Standard deviations

	LIST.PRICE	HOME. TYPE	CITY	BEDS	BATHS	SQFT	LOT.SIZE	YEAR.BUILT	PARKING	DAYS.ON.MKT
mean	1526136.6	2.99	2.24	3.48	2.95	2301.6	14228.24	1982.28	2.73	95.28
standard deviation	1820457.2	0.96	0.77	1.4	1.26	1198.4	37877.53	23.75	1.97	125.11

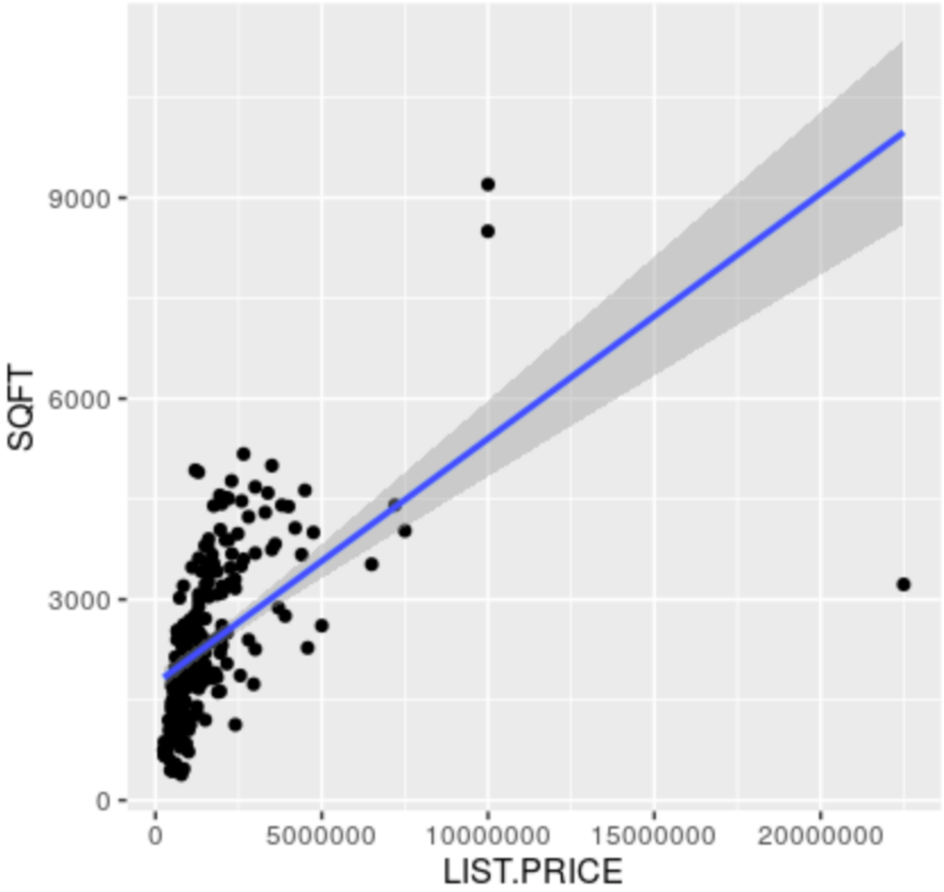
Correlation Matrix



	LIST.PRICE	HOME.TYPE	CITY	BEDS	BATHS	SQFT	LOT.SIZE	YEAR.BUILT	PARKING	DAYS.ON.MK
LIST.PRICE	1	0.02	-0.17	0.29	0.37	0.56	-0.14	-0.02	0.19	0.2
HOME.TYPE	0.02	1	-0.03	0.11	0.16	0.12	-0.47	0.29	-0.15	-0.1
CITY	-0.17	-0.03	1	-0.02	-0.08	-0.07	0.22	-0.04	-0.02	-0.08
BEDS	0.29	0.11	-0.02	1	0.78	0.72	-0.34	0.11	0.45	0.05
BATHS	0.37	0.16	-0.08	0.78	1	0.84	-0.27	0.4	0.41	0.01
SQFT	0.56	0.12	-0.07	0.72	0.84	1	-0.25	0.34	0.39	-0.05
LOT.SIZE	-0.14	-0.47	0.22	-0.34	-0.27	-0.25	1	-0.04	-0.15	-0.06
YEAR.BUILT	-0.02	0.29	-0.04	0.11	0.4	0.34	-0.04	1	-0.16	-0.05
PARKING	0.19	-0.15	-0.02	0.45	0.41	0.39	-0.15	-0.16	1	0.05
DAYS.ON.MK	0.2	-0.1	-0.08	0.05	0.01	-0.05	-0.06	-0.05	0.05	1

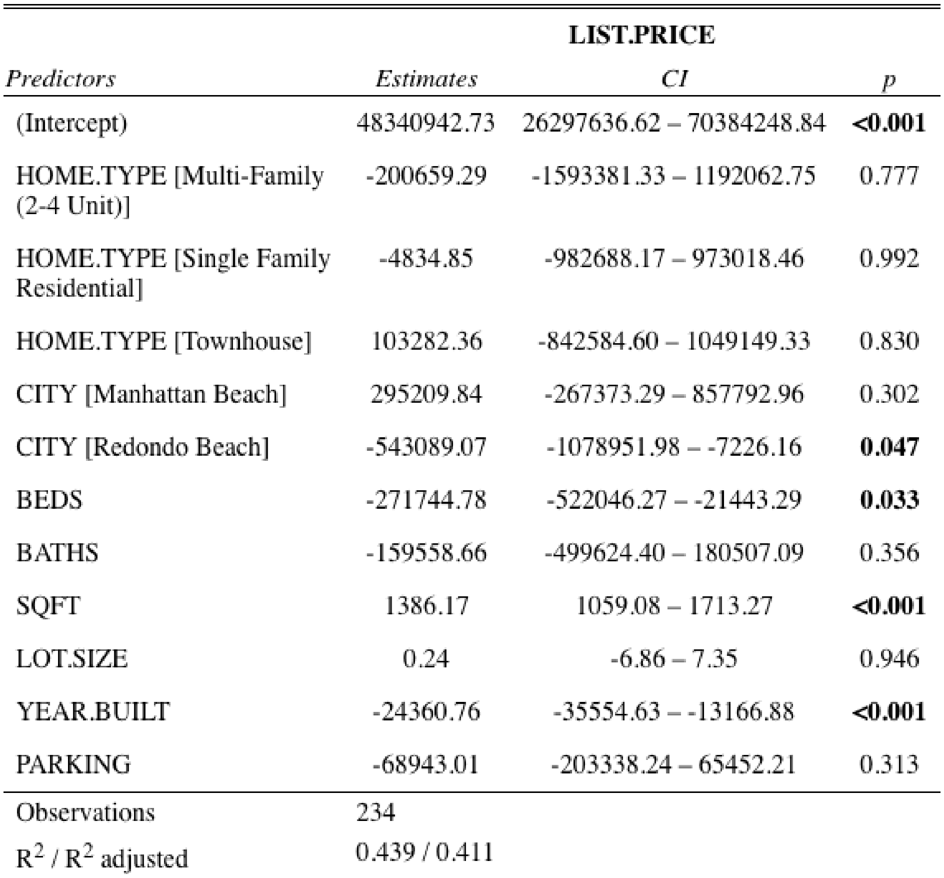
From the above correlation matrix, we can see again that List Price is most strongly correlated with Size (Square Feet). What this tells us that we didn't already know, is that next to size, number of bathrooms and number of bedrooms are closely correlated, relative to the other variables in the data set. This indicates that square footage, number of bathrooms, and number of bedrooms will be important to include in our regression model.

The below scatterplot that was produced from our initial full data multiple linear regression formula, is also one that shows the strong positive correlation between List Price and Size.



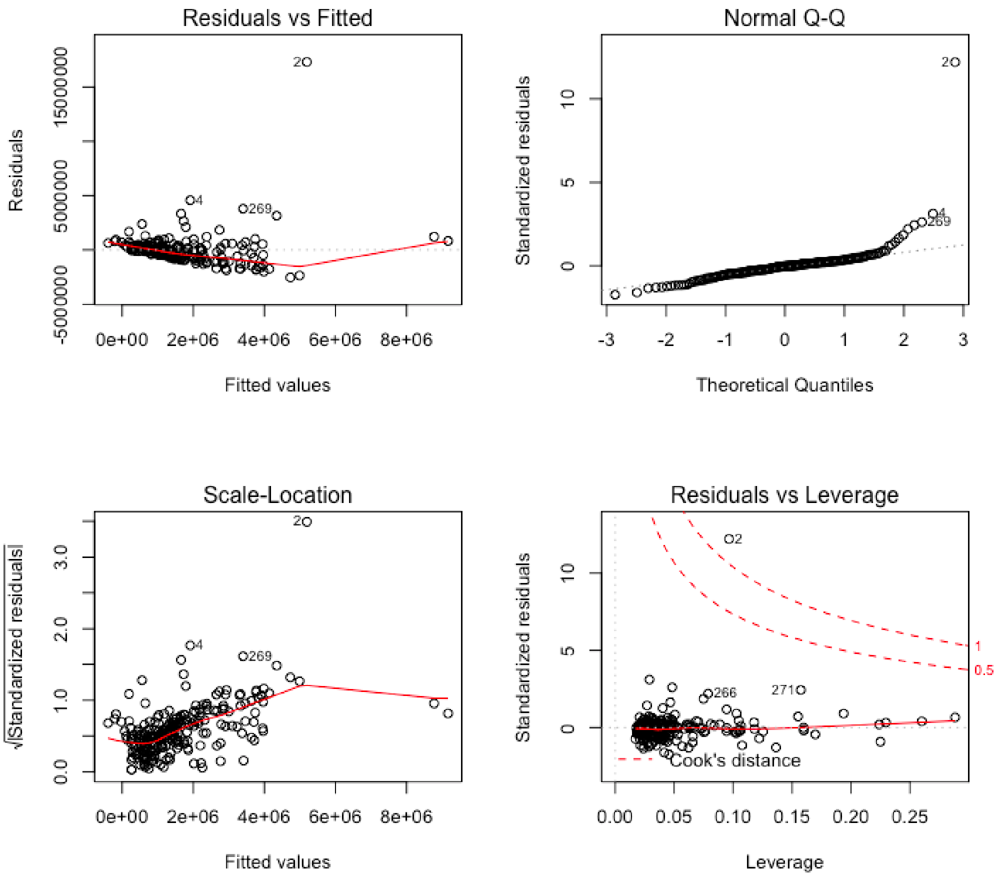
V. **Methods:**

I began by fitting a multiple linear regression model including all variables except Days On Market. The predicted List Price produced by this model should show the price for the property estimated with only information about the property itself, not the Days On Market.



		LIST.PRICE	
Predictors	*Estimates*	*CI*	*p*
(Intercept)	48340942.73	26297636.62 – 70384248.84	**<0.001**
HOME.TYPE [Multi-Family (2-4 Unit)]	-200659.29	-1593381.33 – 1192062.75	0.777
HOME.TYPE [Single Family Residential]	-4834.85	-982688.17 – 973018.46	0.992
HOME.TYPE [Townhouse]	103282.36	-842584.60 – 1049149.33	0.830
CITY [Manhattan Beach]	295209.84	-267373.29 – 857792.96	0.302
CITY [Redondo Beach]	-543089.07	-1078951.98 – -7226.16	**0.047**
BEDS	-271744.78	-522046.27 – -21443.29	**0.033**
BATHS	-159558.66	-499624.40 – 180507.09	0.356
SQFT	1386.17	1059.08 – 1713.27	**<0.001**
LOT.SIZE	0.24	-6.86 – 7.35	0.946
YEAR.BUILT	-24360.76	-35554.63 – -13166.88	**<0.001**
PARKING	-68943.01	-203338.24 – 65452.21	0.313
Observations	234		
R^2 / R^2 adjusted	0.439 / 0.411		

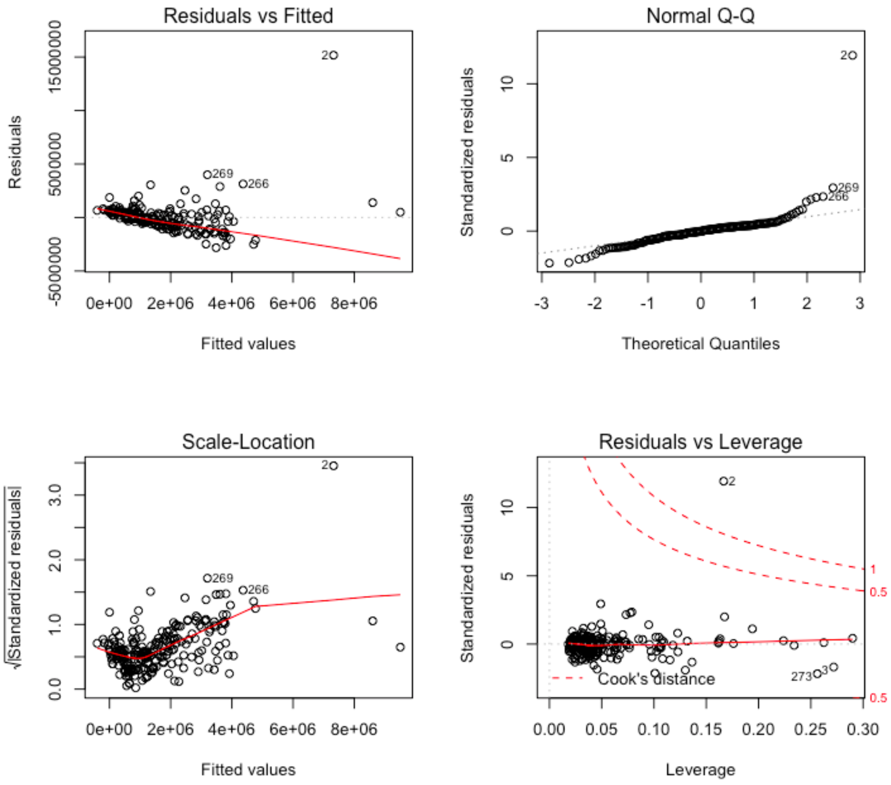
This model shows that the most significant predictors of a houses List Price are being in Redondo Beach, the number of Bedrooms, the Square Footage, and the Year the house was built. Interestingly, the coefficients for the number of bedrooms and bathrooms are negative. This means that after controlling for all other variables (except for Days On Market), fewer bedrooms and bathrooms are associated with higher price of house. Unsurprisingly, when running single linear regression models with only bedrooms and only bathrooms respectively as predictors, the coefficients are positive. This means that controlling for the other variables leads to a negative correlation between beds/baths and price. The R^2 on this model is .439, which means that the model explains about 43.9% of the variability in List Price.

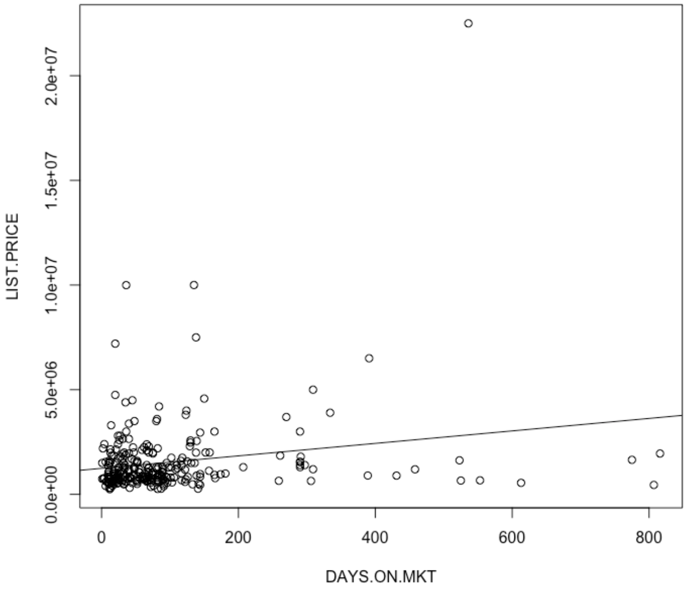


In the diagnostic plots for the model containing all predictors except Days On Market, we see that there are a few outliers that may be affecting our model.

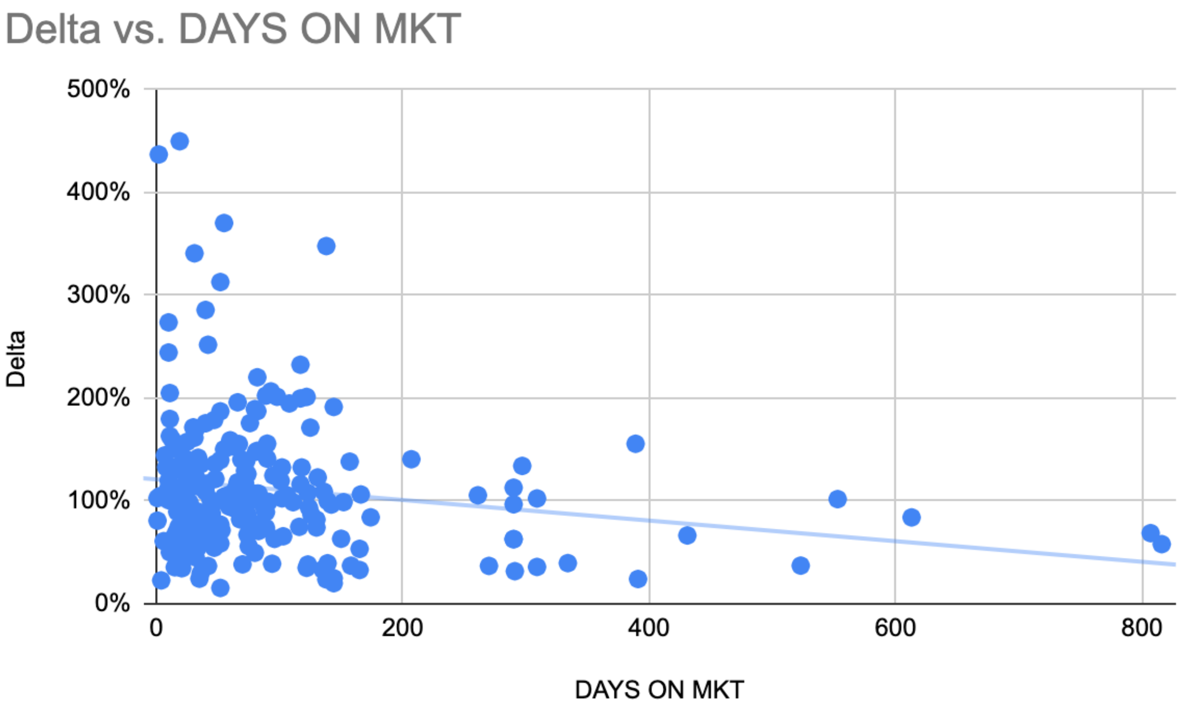
To continue to test my hypothesis, I ran another regression model with all predictors, including Days On Market. This model showed that higher Days On Market were associated with a higher list price at a statistically significant level after controlling for all other variables. This supports my idea that houses that are priced higher take longer to sell. However, it does not tell us whether that is simply because the house has a high price, or because the house is overpriced from what it should be.

	LIST.PRICE		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	37701033.91	16781834.99 – 58620232.84	**<0.001**
HOME.TYPE [Multi-Family (2-4 Unit)]	385491.68	-931405.34 – 1702388.70	0.565
HOME.TYPE [Single Family Residential]	423806.89	-501716.51 – 1349330.29	0.368
HOME.TYPE [Townhouse]	443087.42	-448466.44 – 1334641.28	0.328
CITY [Manhattan Beach]	287695.01	-238065.52 – 813455.54	0.282
CITY [Redondo Beach]	-532830.47	-1033626.01 – -32034.94	**0.037**
BEDS	-317451.31	-551888.84 – -83013.79	**0.008**
BATHS	-230357.96	-549082.81 – 88366.88	0.156
SQFT	1440.00	1133.76 – 1746.23	**<0.001**
LOT.SIZE	2.55	-4.14 – 9.24	0.454
YEAR.BUILT	-19280.44	-29884.83 – -8676.04	**<0.001**
PARKING	-80206.97	-205863.23 – 45449.29	0.210
DAYS.ON.MKT	5064.36	3332.30 – 6796.42	**<0.001**
Observations	234		
R^2 / R^2 adjusted	0.513 / 0.486		





To continue to test my hypothesis, I calculated a value summarizing the difference between the predicted price and the actual list price. I did this by dividing the predicted price by the actual list price for each house. I named this value "Delta." I expected that this value would increase as Days On Market increased, because an overpriced house would be less likely to sell quickly than a correctly priced house.



When I plotted the delta value against the Days On Market, I found that in fact, the trendline was negative. This means that as the Days On Market increased, the Delta value actually decreased.

What this tells us, is that something is missing in the model, the analysis, or the data. Having laboriously sifted through every amount of information I could find on manipulating the data, and also on running the data in various models using every combination of variables that was possible and still finding a similar result, I am confident that the issue was not in the model but in the data. As I referenced in the introduction and in other sections of this study, we have already learned that certain data is missing: We know that distance from the beach is necessary. We also assumed, which the study so far supports, is that the information on interior materials is also needed to correctly determine price accuracy before making the Delta equation. Subsequently, my revised hypothesis is that the conclusion of correlating Delta with Days On Market CAN be done, albeit the answer lies within collecting this additional data.

VI. Conclusion:

After running roughly a dozen models from different perspectives, I have ultimately found that there is not a strong correlation, based on this data, between what the price was set at by the agent, and what the price should have been.

This means the following: I believed that it would make sense for there to be (1) What the list price is set at according to the seller or real estate agent at the time, (2) What the list price *should* have been according to what we know about the property from the data, and (3) A correlation between the Delta of how far away the correct List Price is from the agent estimated List Price, and the Days On Market. In other words, a correlation would be made between the accuracy of the agent's pricing and how fast the property sells.

After running many models to try and properly determine what the price should have been, and then take the expected correct price and run it against what the price was determined to be by the agent, it would have made sense that the amount of variation between the estimated price and the correct price would have been the determining factor in how long the house was on the market. Unfortunately, this theory did not hold true when analyzing the data, and I am therefore led to believe that there is foundational data missing that would allow me to correctly determine what the appropriate Price should have been for a house to sell it in the least time possible.

While I was originally discouraged that I wasn't able to see a simple and obvious linear correlation between what would essentially be a "mis-pricing", which is logically what should be the case (Over-Pricing = longer Days On Market), what this data showed, based only on this data, is that since you can not draw that correlation, that must mean there are missing variables.

As I stated earlier, I happen to know as a real estate professional, there is indeed great value in the details of the house, may that be the construction, the finishes, the materials, the style, the general condition, etc. Additionally, in this data set, there is other important information missing such as if the house has a pool or not.

While I hope the variables provided were enough to at least draw a general correlation that could be tweaked by incorporating further value factors such as the quality of how the house is built, when it was updated, etc, we have now learned this is not the case. The subsequent conclusion of learning a correlation can not be drawn only on basic metrics such as Size, Rooms, Year Built etc, is that there is much more value in how the house is constructed, designed, and things such as the quality of materials than one might assume.

This is, although not a black and white picture, very important for a seller or an agent to understand and study. Perhaps a deeper data set that incorporates the expenses of materials used, the style similarity or difference with other houses in the neighborhood, and like factors, which are not so easily pulled from databases like Realtor.com or Zilliow.com, would help us understand more about how to get proper price-per-square-foot values for houses.

VII. <u>Next Steps:</u>

As an action item to be taken away from this study, is that a following study would need to be done in order to reach a true conclusion to ultimately support or reject the hypothesis:

Further data to be collected on each of the houses in our data set:

1. Interior materials used with associated costs
2. Distance from the ocean
3. If the house contains a pool
4. Sale price

Following study process:

We may use the Distance From Ocean data to further break down and separate out individual markets so that similar market comparisons can be made to give accurate pricing. Then use the Interior Material Costs to further normalize the houses in a data set in order to find correct pricing. Assign a general value of how much pools add to a house, and create a yes/no scenario to account for that pool value. Incorporate Sale Price. Then run a multiple linear regression on the new data, after substituting Sale Price with List Price, using Sale Price as the dependent variable, and subtract Days On Market, in order to determine how the houses should be priced according to the data, to result in “Correct Price”

Next, create a new “Delta” column showing the difference between the Correct Price and the List Price. Finally, using this Delta, sort according to Days On Market, and run a linear regression between Delta & Days On Market in order to draw a correlation between the mispricing by the agent & time to sell. This should supply the significance of correct pricing, along with allowing us the ability to determine Correct Pricing going forward, and of course, then hopefully obey it!