Hemen Asfaw

BSAN 6400

Professor Tao

7 December 2024

# Modeling Housing Prices in Los Angeles and Miami Using Redfin Data

## *Scenario*

*You are continuing your role as a consultant for Ms. Jones, who now seeks insights into factors influencing real estate prices in Los Angeles and Miami. You have already gathered the relevant data, and insights from your previous project have provided Ms. Jones with an understanding of real estate price distributions across areas of interest. Now, she aims to develop a predictive model to forecast house prices and price per square foot (price/sqft) based on key property features to inform her real estate investment strategy. Note that price/sqft cannot be used to predict house price, and vice versa, since these values should be unknown when making predictions.*

## Project Introduction

The dataset, originally scraped from the Redfin real estate website and later organized in Excel, was carefully cleaned and prepared for analysis prior to modeling. The first section of this paper addresses descriptive analytics questions, offering a clear overview of the dataset's characteristics before building predictive models.

Next, we begin with simple linear regression models for both **house price** and **price per square foot (price/sqft)**. Starting with simple regression allows us to examine the individual relationship between a single predictor (such as square footage or lot size) and the target variable. This step is valuable because it provides interpretability, highlights the direct influence of a single feature, and establishes a baseline performance for comparison.

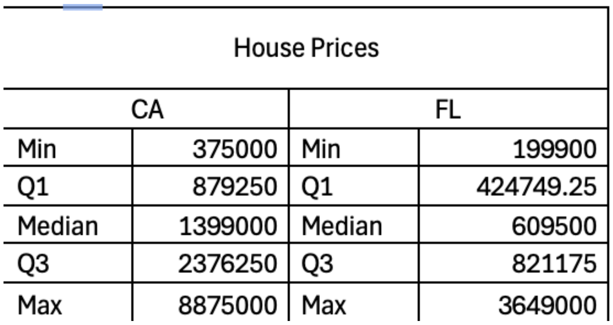
Following this, we extend the analysis to multiple linear regression. By incorporating several property features simultaneously (e.g., number of bedrooms, bathrooms, location indicators), multiple regression better captures the complexity of real estate pricing and produces more accurate predictions than simple regression alone.

Finally, the paper concludes with a discussion of the model results and a set of recommendations that can guide Ms. Jones in refining her real estate investment strategy.

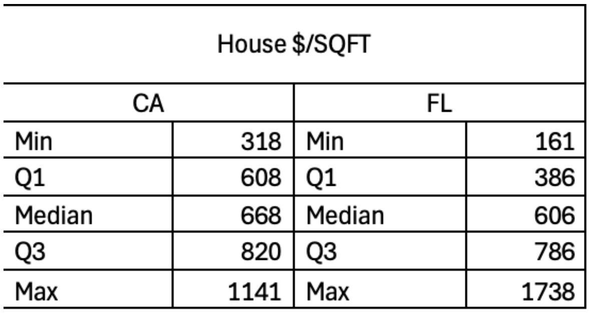
## Data Exploration

Before building predictive models, it is important to explore the dataset to understand its overall structure and the relationships among key variables.

### Summary Statistics



| House Prices | | | |
|---|---|---|---|
| CA | | FL | |
| Min | 375000 | Min | 199900 |
| Q1 | 879250 | Q1 | 424749.25 |
| Median | 1399000 | Median | 609500 |
| Q3 | 2376250 | Q3 | 821175 |
| Max | 8875000 | Max | 3649000 |



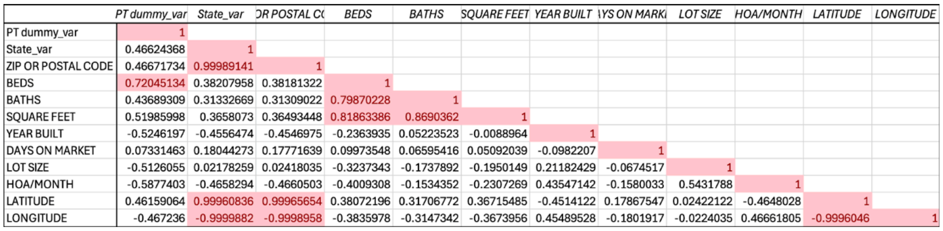
| House $/SQFT | | | |
|---|---|---|---|
| CA | | FL | |
| Min | 318 | Min | 161 |
| Q1 | 608 | Q1 | 386 |
| Median | 668 | Median | 606 |
| Q3 | 820 | Q3 | 786 |
| Max | 1141 | Max | 1738 |

The tables above present descriptive statistics for house prices and price per square foot (price/sqft) in California (CA) and Florida (FL). These summaries provide a sense of the distribution of values in each market.

- House Prices: California homes generally have higher values across all quartiles compared to Florida, with the median price in CA nearly double that of FL. The wide range between minimum and maximum values highlights substantial variability in both markets.
- House Price per Square Foot: While median price/sqft values are relatively close between the two states (668 in CA vs. 606 in FL), Florida exhibits a higher maximum value, suggesting the presence of luxury properties or highly localized pricing dynamics.

These statistics help establish the baseline differences in affordability and valuation between the two regions.

### Correlation Analysis



| | PT dummy_var | State_var | OR POSTAL C( | BEDS | BATHS | SQUARE FEET | YEAR BUILT | \YS ON MARK | LOT SIZE | HOA/MONTH | LATITUDE | LONGITUDE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PT dummy_var | 1 | | | | | | | | | | | |
| State_var | 0.46624368 | 1 | | | | | | | | | | |
| ZIP OR POSTAL CODE | 0.46671734 | 0.99989141 | 1 | | | | | | | | | |
| BEDS | 0.72045134 | 0.38207958 | 0.38181322 | 1 | | | | | | | | |
| BATHS | 0.43689309 | 0.31332669 | 0.31309022 | 0.79870228 | 1 | | | | | | | |
| SQUARE FEET | 0.51985998 | 0.3658073 | 0.36493448 | 0.81863386 | 0.8690362 | 1 | | | | | | |
| YEAR BUILT | -0.5246197 | -0.4556474 | -0.4546975 | -0.2363935 | 0.05223523 | -0.0088964 | 1 | | | | | |
| DAYS ON MARKET | 0.07331463 | 0.18044273 | 0.17771639 | 0.09973548 | 0.06595416 | 0.05092039 | -0.0982207 | 1 | | | | |
| LOT SIZE | -0.5126055 | 0.02178259 | 0.02418035 | -0.3237343 | -0.1737892 | -0.1950149 | 0.21182429 | -0.0674517 | 1 | | | |
| HOA/MONTH | -0.5877403 | -0.4658294 | -0.4660503 | -0.4009308 | -0.1534352 | -0.2307269 | 0.43547142 | -0.1580033 | 0.5431788 | 1 | | |
| LATITUDE | 0.46159064 | 0.99960836 | 0.99965654 | 0.38072196 | 0.31706772 | 0.36715485 | -0.4514122 | 0.17867547 | 0.02422122 | -0.4648028 | 1 | |
| LONGITUDE | -0.467236 | -0.9999882 | -0.9998958 | -0.3835978 | -0.3147342 | -0.3673956 | 0.45489528 | -0.1801917 | -0.0224035 | 0.46661805 | -0.9996046 | 1 |

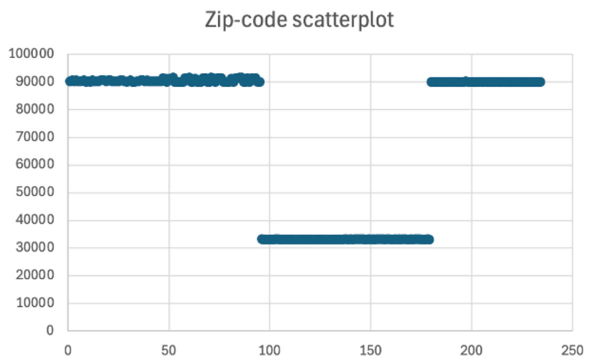
A correlation matrix was computed to assess relationships between independent variables. To make interpretation clearer, correlations greater than or equal to |0.7| were flagged in red.

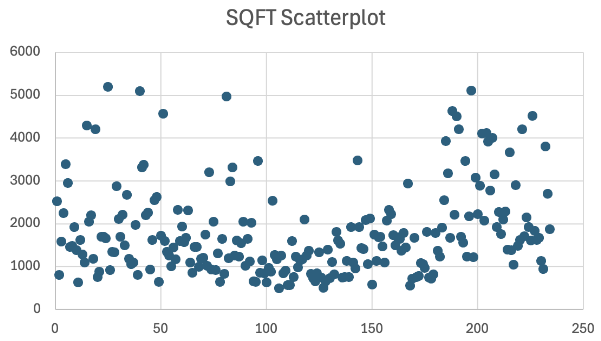
- Strong correlations exist between beds, baths, and square footage, which is expected since larger homes typically have more rooms.

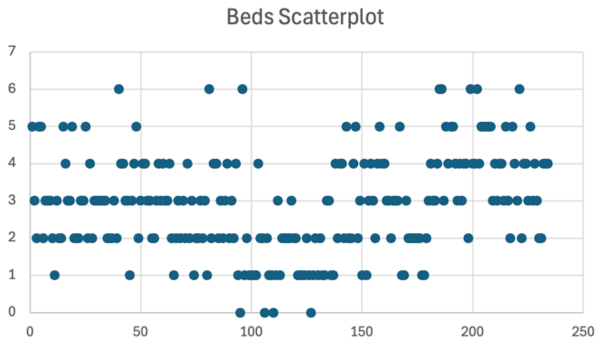
- Latitude and longitude are almost perfectly (negatively) correlated, reflecting geographic coding in the dataset rather than meaningful predictive differences.
- High correlations among certain features raise the risk of multicollinearity in regression models, which must be addressed when selecting predictors for multiple regression.

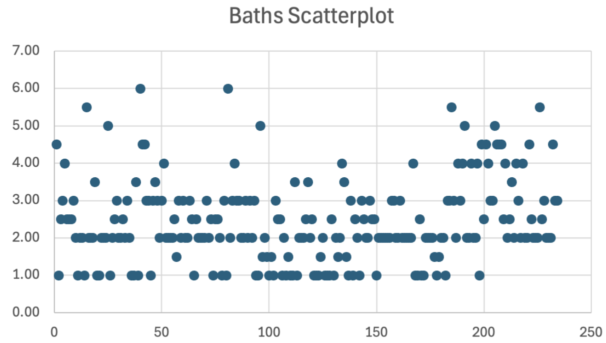
This step was critical for identifying redundant predictors and avoiding distorted model coefficients.

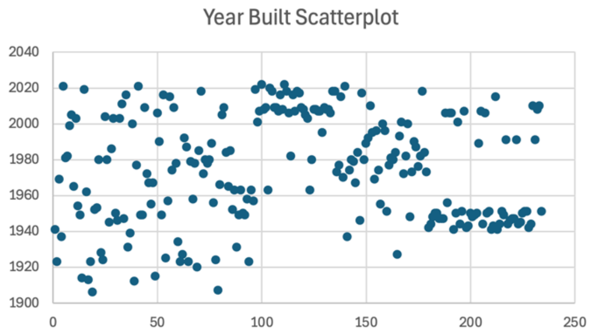
Scatterplots

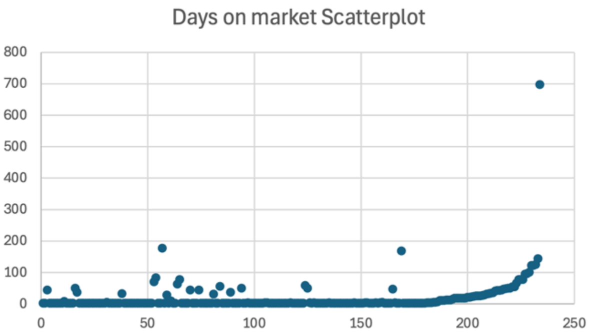


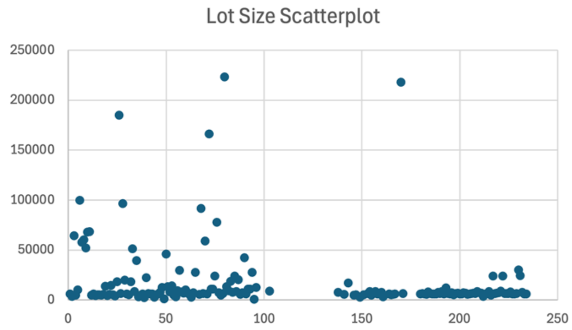


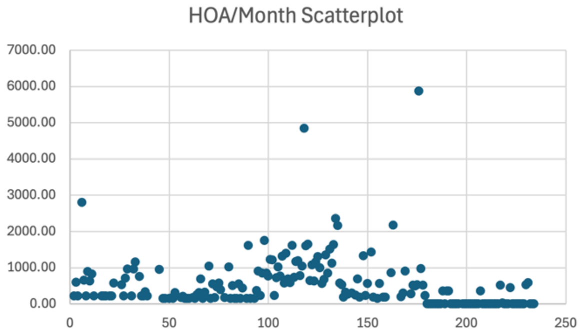


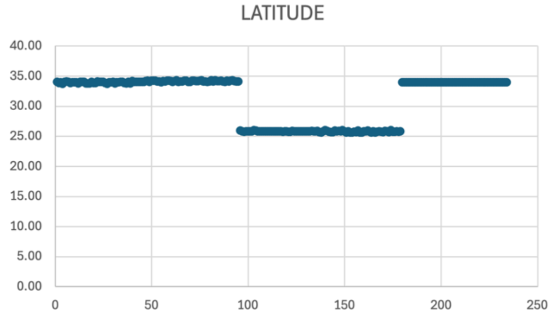


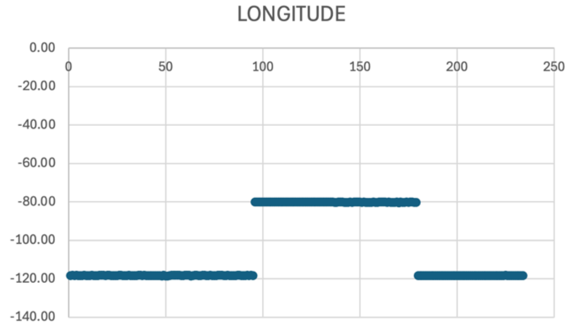












Scatterplots for each variable were generated to visualize distributions and potential trends.

- Zip Code: The distribution is clustered around a few ranges, reflecting the localized nature of the dataset (LA and Miami markets).
- Square Footage vs. Price: Considerable spread exists, but the positive relationship suggests sqft is a strong predictor of house price.
- Beds and Baths: These are mostly concentrated between 2–5, typical for residential properties. Outliers (e.g., 6+ bedrooms) may influence model performance.
- Year Built: The data spans over a century, with many properties built after 1950. More recent builds may be associated with higher valuations.
- Days on Market: While most homes sell relatively quickly, some extreme values suggest properties that linger on the market, potentially influencing price adjustments.
- Lot Size: While many properties cluster at relatively modest lot sizes, there are significant outliers with exceptionally large parcels of land. This wide spread may introduce skewness into the data, making transformations or robust regression techniques worth considering.
- HOA/Month: Most homes either have low or zero HOA fees, but a subset of properties includes very high HOA costs (over $5,000). These outliers could disproportionately influence predictions if not handled carefully.
- Latitude and Longitude: The geographic scatterplots clearly distinguish the two markets—Los Angeles and Miami—through separate latitude and longitude clusters. These variables are critical in separating regional pricing patterns but, given their perfect correlation, they may need to be encoded differently (e.g., as a state/city indicator) to avoid multicollinearity.

Together, these visualizations confirm that the dataset is varied, contains meaningful predictors, and highlights the importance of handling multicollinearity when transitioning to regression analysis.

## Simple Linear Regression

We first developed simple linear regression models to establish baseline relationships between property size (square footage) and the two dependent variables: house price and price per square foot (price/sqft).

House Price Model:

$$\text{House Price} = -265{,}018.1 + 936.48 \times \text{Sqft}$$

Price per Sqft Model:

$$\text{Price/Sqft} = 639.75 + 0.062 \times \text{Sqft}$$

## Model Analysis

*Model A (House Price)*

This model uses square footage as the sole predictor of house price. The model achieved an $R$^2 value of 65.47%, meaning square footage explains a substantial portion of the variation in house price. However, about 34.63% of the variation remains unexplained. The significance F value was below 0.05, and the p-value for the square footage coefficient was also less than 0.05, confirming the model's statistical significance. While this is a reasonably strong model, the unexplained variance suggests that house price is influenced by additional factors (such as location, lot size, or number of bedrooms/bathrooms) that should be incorporated into a multiple regression framework.

*Model B (Price/Sqft)*

This model attempts to predict price per square foot using square footage. It produced an $R$^2 value of only 4%, leaving 96% of the variation unexplained. Although the F-statistic and coefficient p-value were statistically significant, the predictive power is very weak. This indicates that square footage alone does not meaningfully predict price per square foot, which makes sense because price/sqft is more influenced by location, amenities, and market conditions than by raw size.

Starting with simple regression was valuable because it highlighted the strong direct relationship between square footage and overall house price (Model A), while also showing the limitations of using square footage to predict price/sqft (Model B). These insights motivate the transition to multiple linear regression, where additional explanatory variables can be included to capture more of the complexity underlying housing prices.

## Multiple Linear Regression Analysis

To improve predictive accuracy beyond simple regression, I developed multiple linear regression models for both **house price** and **price per square foot (price/sqft)**.

### Step 1: Data Transformation and Robust Regression

House prices were right-skewed, so I applied a log transformation to normalize the distribution, which improved model performance. I also tested robust regression to address issues of outliers and heteroskedasticity, resulting in higher R2 values compared to ordinary least squares.

### Step 2: Variable Selection

I used forward and backward selection, both of which identified the same predictors: square feet, state, zip code, property type, baths, and HOA/month. However:

- Zip code and state were highly correlated, so zip code was removed.
- HOA/month was excluded due to extensive missing values.

This left state and square feet as the most reliable predictors, yielding an adjusted R2 of 66.88%.

### Step 3: Full Model Reassessment

Unsatisfied with the limited model, I re-ran the regression including all available variables. While this improved raw accuracy, many predictors were statistically insignificant, and several exhibited high variance inflation factors (VIFs) due to collinearity—especially among location-based variables (latitude, longitude, zip code, state).

To resolve this, I:

- Retained state as the primary location indicator.
- Excluded lot size and HOA/month (again, due to missing values).
- Gradually removed statistically insignificant predictors ($p>0.05$).

This refinement produced a model with state, beds, baths, and square feet, achieving an adjusted R2 of 67.6%. Although multicollinearity among beds, baths, and square feet was a concern, VIF values were mostly below 5, suggesting tolerable levels.

### Step 4: Splitting by Property Type

To address lingering concerns, I split the dataset into Single-Family Residences (SFRs) and Townhouses. This allowed me to include variables that were more meaningful for each property type:

- **SFR Model:** State, year built, square feet, lot size (adjusted R2=69.5%).
- **Townhouse Model:** State, year built, square feet, HOA/month (adjusted R2=77%).

Both split models outperformed the joint model in accuracy and interpretability, making them better suited for forecasting house prices.

### Step 5: Price per Square Foot Models

A similar approach was taken for predicting price/sqft:

- The joint model (state, beds, baths) explained only 22.11% of variation, showing weak predictive power.
- Splitting by property type significantly improved results:
  - SFR Model: State, year built, square feet, lot size (R2=37.42%%).
  - Townhouse Model: State, year built, beds, square feet, HOA/month (R2=58.21%).

### Final Models

*House Price (log-transformed)*

- **SFR:**
  $\text{Log(Price)} = 18.2186 + 0.5780(\text{State}) - 0.0028(\text{Year Built}) + 0.0005(\text{Sqft}) + 0.00000572(\text{Lot Size})$
- **Townhouse:**
  $\text{Log(Price)} = -15.2985 + 0.5786(\text{State}) + 0.0107(\text{YearBuilt}) + 0.0004(\text{Sqft}) + 0.0003(\text{HOA/Month})$

*Price/Sqft (log-transformed)*

- **SFR:**
  $\text{Log(Price/Sqft)} = 12.03 + 0.6012(\text{State}) - 0.0031(\text{Year Built}) + 0.0000465(\text{Sqft}) + 0.00000559(\text{Lot Size})$
- **Townhouse:**
  $\text{Log(Price/Sqft)} = -15.2985 + 0.5682(\text{State}) - 0.1104(\text{Beds}) - 0.0002(\text{Sqft}) + 0.0109(\text{Year Built}) + 0.0003(\text{HOA/Month})$

### Interpretation and Takeaways

Because the dependent variables are log-transformed, coefficients represent percentage changes rather than absolute changes. For instance, the state coefficient (~0.57) indicates that a change in state is associated with roughly a 57.8% change in house price, holding other variables constant.

Splitting the dataset by property type produced the most effective models, balancing statistical significance, interpretability, and predictive accuracy. These models provide Ms. Jones with more actionable insights, as they account for differences between SFRs and townhouses while remaining robust and relatively simple.

## Final Analysis and Conclusions

### Brief Analysis

From the price models, we observe that state and year built are the strongest predictors of property values across both single-family residences (SFRs) and townhouses.

- State: A change in state results in approximately a 57.8% increase in SFR prices and a 57.86% increase in townhouse prices, highlighting the critical role of location in property valuation.
- Year Built: For townhouses, newer construction leads to higher prices, with a 1.07% increase per year built. For SFRs, the relationship is slightly negative, with a 0.28% decrease per year built, which may be due to older SFRs often being larger, more unique, or situated on bigger lots.
- Square Footage and Lot Size: In SFRs, square footage and lot size contribute to higher prices (0.05% and 0.000572%, respectively), though their effects are small relative to location and age. In townhouses, square footage and HOA/month both play a role, contributing 0.04% and 0.03% to price increases, respectively.

For price per square foot (price/SQFT), similar trends emerge:

- SFRs: State and year built again dominate, with a 60.12% increase due to state and a 0.31% decrease per year built. Square footage and lot size provide smaller contributions (0.00465% and 0.000559%).
- Townhouses: State increases price/SQFT by 56.82%, while the number of bedrooms has a negative effect, reducing price/SQFT by 11.04% per additional bedroom. Square footage also has a negative effect, while newer construction and higher HOA fees increase price/SQFT.

These results align with real estate logic: as homes become larger (through added bedrooms or increased square footage), the total cost rises but the **price per square foot falls**. This reflects diminishing returns to scale and how additional space dilutes per-unit cost.

## Recommendations

For realtors and investors, the models provide several actionable insights:

1. Location matters most: Properties in California command significantly higher total prices and higher price per square foot compared to Florida. Location should remain the first consideration in pricing strategy.
2. Year built influences differently by property type:
   - Townhouses: Newer builds tend to be more valuable.
   - SFRs: Older homes can sometimes be more valuable, possibly due to larger lot sizes or historic appeal.
3. Smaller homes vs. larger homes:
   - Buyers seeking smaller homes should focus on areas with higher price/SQFT, which typically reflects strong demand for compact, efficient living spaces.
   - Buyers prioritizing spaciousness (more bedrooms, larger square footage) may benefit from lower price/SQFT properties, especially in the townhouse market.
4. Lot size and HOA/Month add value:
   - Larger lot sizes increase SFR value.
   - Higher HOA/month costs correlate with higher townhouse value, likely because higher fees are associated with more amenities or desirable communities.

## Limitations and Future Improvements

These models are useful but not perfect. They capture general relationships but leave room for improvement:

- Predictor coverage: Including additional features such as proximity to schools, access to amenities, and neighborhood demographics could enhance explanatory power.
- Macroeconomic factors: Housing prices are strongly influenced by external elements like inflation, interest rates, and lending policies, which were not included here.
- Model accuracy: While adjusted R2R2 values of 67–77% are strong for house price, the models for price/SQFT perform less well, especially for joint datasets. Splitting by property type improved interpretability, but further refinements could still be made.

Overall, the analysis confirms that location, property age, and size are the strongest drivers of housing values. Splitting models by property type (SFR vs. townhouse) significantly improves performance, balancing statistical rigor with practical interpretability. These insights provide Ms. Jones with a strong foundation for evaluating investment opportunities in the Los Angeles and Miami markets.