Analysis of the Houston Real Estate Market

Brendan Stegner

Business Understanding

In the real estate industry, the ability to accurately predict prices is one of the most important in fast-changing and heterogeneous markets such as that of Houston. Traditionally, pricing has been done by real estate professionals based on experience, neighborhood knowledge, and comparative market analysis. While these methods are still relevant, the move towards data-driven decision making provides a way to more accurately and efficiently model large datasets. This project is to apply statistical techniques to Houston residential real estate data to identify important price determinants, clean the data, and get it ready for modeling.

This project looks at how different attributes, structural attributes such as the number of bedrooms, number of bathrooms, and the size of the house, location attributes such as zip code and neighborhood name, etc., relate to the dependent variable, PRICE. While preparing the data, we must face such problems as missing values, inconsistent modeling types, outliers, and high-cardinality categorical variables. This is important for the sake of predictability and model reliability. The result is a comprehensive preparation pipeline, ready for advanced modeling.

The 3 V's of Big Data

In this analysis, the Houston real estate dataset is a good example of big data, which has three main dimensions: Volume, Velocity, and Variety.

Volume: The dataset has over 6,000 records and dozens of variables that capture a lot of property information. These include attributes such as number of bedrooms and square footage, categorical attributes such as type of property and city, and metadata attributes such as listing status, open house time, and unique identifiers.

Velocity: Real estate pricing is time sensitive. Variables such as PRICE and DAYS ON MARKET are time-dependent and change with the trends of the market, changes in demand, and seasonal trends. It is therefore important to capture and analyze the data in near real time to make better pricing decisions.

Variety: The dataset contains numerical variables (SQUARE FEET, PRICE, etc.), categorical variables (PROPERTY TYPE, CITY, etc.), spatial variables (LATITUDE, LONGITUDE, etc.), temporal fields (OPEN HOUSE START/END TIME, etc.) and even consumer sentiment indicators (FAVORITE, INTERESTED etc.). The URLs further suggest that there is potential to include images or unstructured data, but this was not included in modeling.

Because of this variety, our first challenge is not modeling but cleaning and standardizing—deciding what's useful, what's redundant, and what's missing. This is the basis of everything that is to come.

Data Understanding

The data could not be modeled as it was and needed to be cleaned and reduced in size. Several columns were dropped since they had no variance, were irrelevant to pricing, or had too many missing values. Variables such as STATE, STATUS, FAVORITE, and INTERESTED had the same value for all the rows and did not have any useful information. Fields like SOURCE, MLS, and URL were not related to the home itself and were therefore eliminated.

The variable $/SQUARE FEET was removed because it is a function of PRICE and SQUARE FEET, thus introducing a circular dependency. Temporal variables like SOLD DATE and NEXT OPEN HOUSE START/END TIME were quite incomplete and offered little modeling value. Last but not least, ADDRESS and LAT/LONG were too detailed for modeling purposes but were kept for spatial mapping and visualization (Figure 1).

The next step was to examine the distributions of the core numeric predictors. BEDS, BATHS, and SQUARE FEET had a very strong right skew and some outliers, mainly due to luxury or custom-built homes. These distributions were visualized in Figures 2 through 5, which showed that modeling required imputation, binning, or transformation before modeling.

Analysis and Recoding

One of the most important steps in data preparation was checking the type of model that each predictor was suited for and changing those that were not suitable for linear modeling or were not clear in terms of categories. For instance, PROPERTY TYPE had some non-conventional or rare categories like Vacant Land, Mobile Home, and Multi-Family (5 or more Units). These entries were also a concern since they did not have structural features like BEDS, BATHS, or SQUARE FEET, which would make comparison invalid. Therefore, these rows were dropped and only single-family, condo, and townhouse listings were used.

ZIP received initial treatment as a measurement variable, which was incorrect because ZIP codes function as identifiers instead of measurements. The data was transformed into nominal and then split into 8 categories according to house price averages, which extended from $100K to over $700K (Figure 9). The process decreased the number of categories but maintained relevant geographic separation.

The LOCATION data underwent a similar grouping process as ZIP did, which resulted in 12 distinct groups through price cluster analysis. The ZIP-based analysis gained additional geographic detail through LOCATION, which remained workable for modeling purposes (Figure 10). The use of bins enables the model to fit smoothly while maintaining interpretability by avoiding excessive levels.

JMP's Fit Y by X function was used to study the connection between PRICE and continuous predictors. A strong nonlinear relationship appeared between price and SQUARE FEET because larger homes cost more, but the price growth slows down at higher square footage levels (Figure 6). BEDS and

BATHS showed moderately linear relationships (Figures 7–8), with slight curvature at the high end, suggesting that more than four bedrooms or bathrooms may not provide proportionate returns. The data requires quadratic or spline terms for modeling because linear models fall short in representing the patterns.

Imputation

BEDS, BATHS, and SQUARE FEET had missing values because the dataset contained property types such as vacant land together with incomplete listings. The core real estate pricing variables were essential, so we needed to keep all data rows despite missing values. JMP's distribution-based automated imputation system completed missing data points while keeping the original data distributions intact (Table 1).

The missing data prediction tool generates statistical estimates for missing values through an analysis of observed variable patterns. This technique prevents the introduction of bias and variability reduction that occurs when using mean substitution or regression prediction. The imputation method preserved the natural distribution shapes since the data displays strong skewness as illustrated in Figures 2–5. The imputed values maintained both the central tendency and spread of the variables throughout the dataset.

After imputation, the data contained complete values for BEDS, BATHS, and SQUARE FEET for all listings included in the analysis. The analysis continued with the complete data set without any case reduction, which became crucial for studying specific neighborhoods and higher-priced properties. The visualization showed that Figures 6 through 8 maintained their accuracy and information quality after conducting the imputation process. The cleaned dataset is summarized in Figure 11.

Conclusion

A full data preparation cycle was implemented for modeling property prices in Houston through a residential listing dataset exceeding 6,000 records. The process included eliminating unhelpful variables and improper data types and detecting outliers and properly handling missing values through statistically valid imputation methods. The strategic binning of ZIP and LOCATION variables reduced cardinality but maintained essential location-based differences because they contained price clusters.

The exploratory analysis revealed both linear and nonlinear relationships between core predictors and price. The analysis revealed that SQUARE FEET followed a quadratic pattern whereas BEDS and BATHS exhibited near-linear patterns with minor curvature. Future models should consider adding polynomial terms and employing nonlinear modeling approaches because of these findings.

The cleaned dataset includes PRICE and BEDS, BATHS, SQUARE FEET, GROUP ZIP, and GROUP LOCATION data, which have been binned and imputed for predictive modeling use. This data pipeline is ready to support regression, classification, or machine learning models to help forecast home prices across Houston with clarity, scale, and insight.

References

Lee, Cheng-Few, John C. Lee, and Alice C. Lee. *Statistics for Business and Financial Economics*. 3rd ed., Springer, 2013.

JMP Statistical Software. SAS Institute Inc.

Table 1: Summary of Description of Variables

Number of rows scope	6,638			
Final number of rows	6,542			
Variable	Outliers	Missing	Data Modeling Type	Included
PROPERTY TYPE	0	0	Nominal	Yes
CITY	0	0	Nominal	No
ZIP	0	0	Nominal	Yes
BEDS	0	3	Continuous	Yes
BATHS	3	17	Continuous	Yes
LOCATION	0	4	Nominal	Yes
LOT SIZE	205	1129	Continuous	No
SQ FEET	16	34	Continuous	Yes
DAYS ON MARKET	42	297	Continuous	No
$/SQ FEET	16	34	Continuous	No
LATITUDE	0	0	Continuous	No
LONGITUDE	0	0	Continuous	No
YEAR BUILT	0	348	Continuous	No
PRICE	76	0	Continuous	Yes

Figure 2 – Histogram of Baths

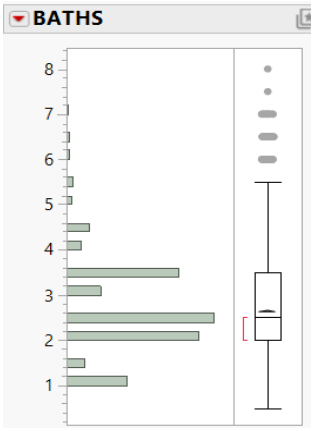


Figure 3 – Histogram of Beds

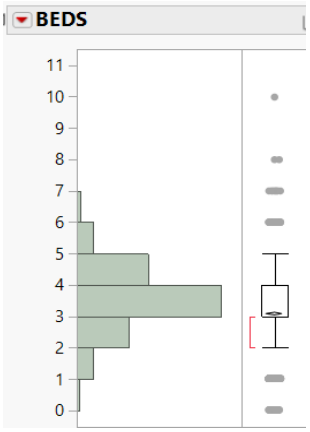


Figure 1: Map of ZIP

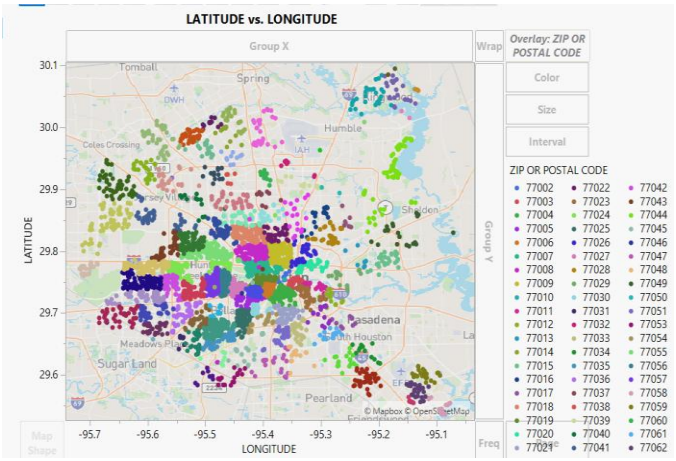


Figure 4 – Histogram of Price

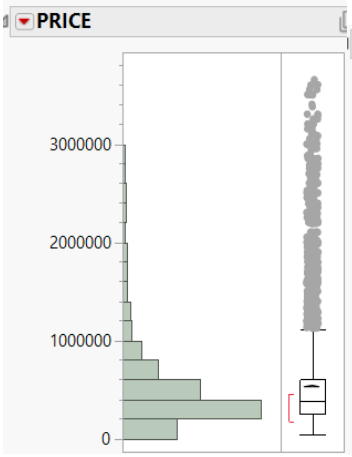


Figure 5 – Histogram of Sq Feet

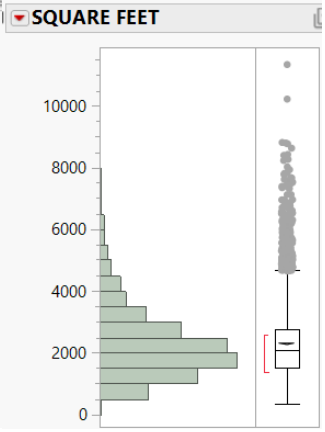


Table 2: Summary of Relationships to Price

Variable	Binned	Relationship to Price	Data Modeling Type
BEDS	No	Linear	Continuous
BATHS	No	Non Linear	Continuous
SQ FEET	No	Non Linear	Continuous
PROPERTY TYPE	No	N/A	Nominal
ZIP	Yes	N/A	Nominal
LOCATION	Yes	N/A	Nominal

Figure 6 – Price by Sq Feet

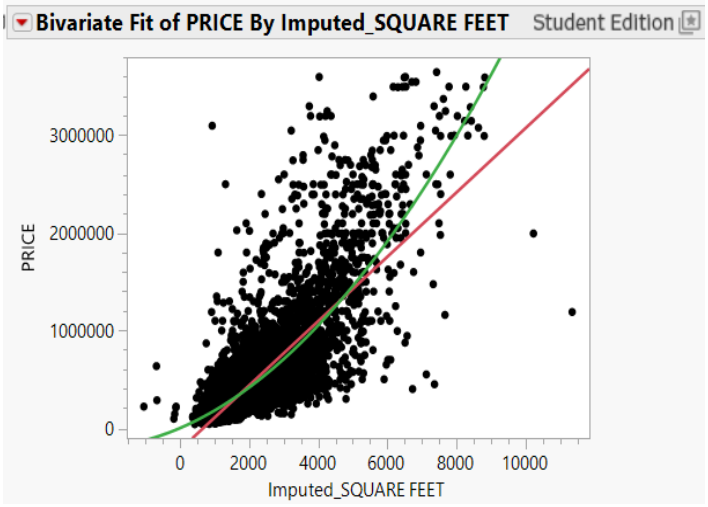


Figure 7: Price by Beds

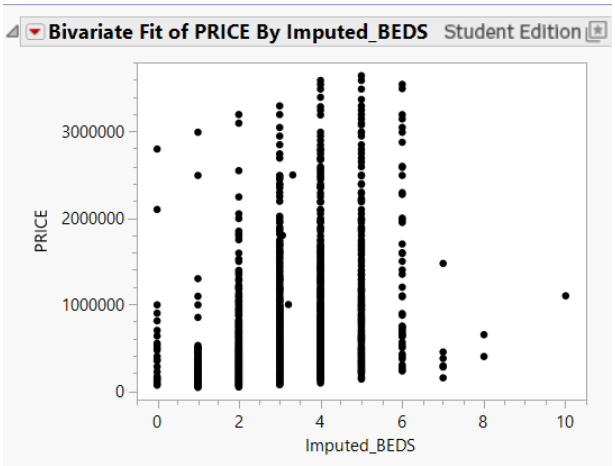


Figure 8: Price by Baths

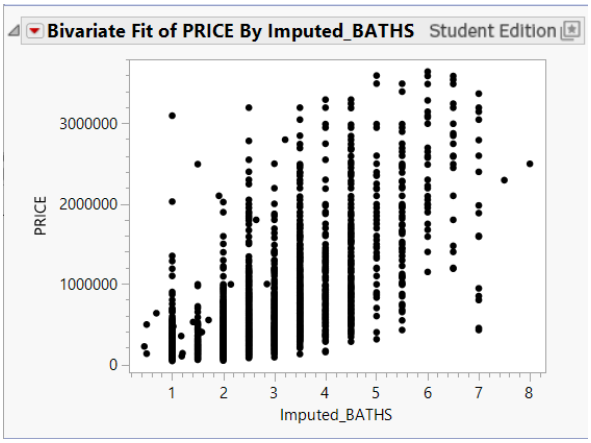


Figure 9: Map of Properties by ZIP

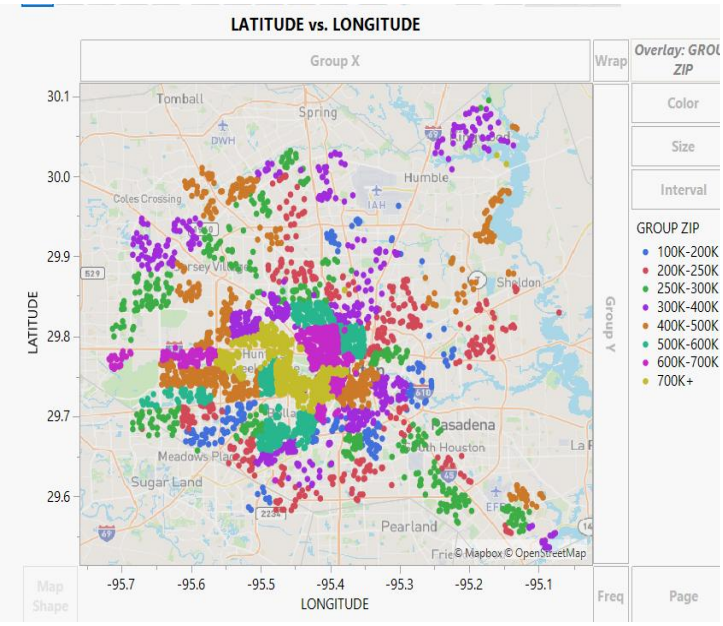


Figure 10: Map of Properties by Location

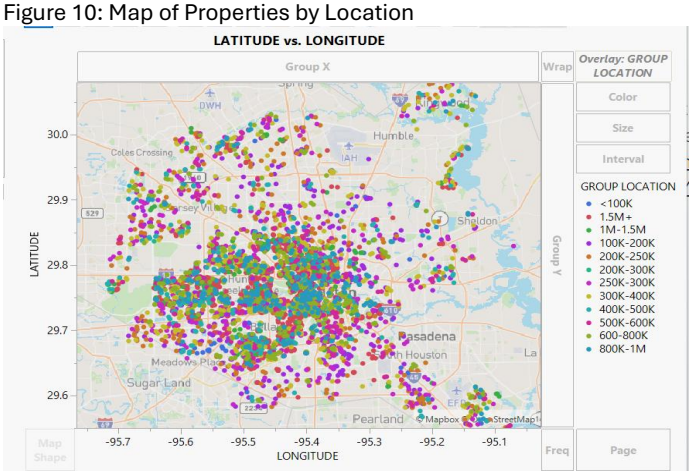
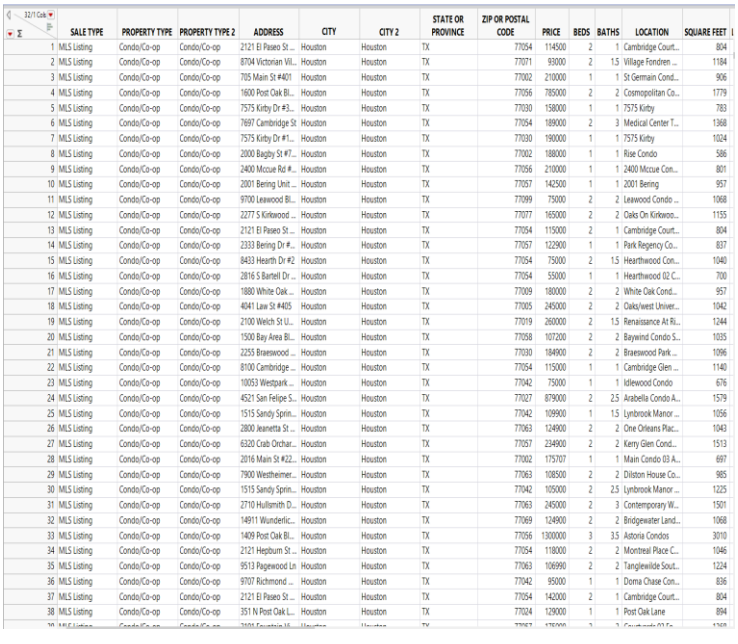


Table 3: Number of Properties in each Bin

GROUP LOCATION	N
<100K	76
1.5M+	506
1M-1.5M	261
100K-200K	407
200K-250K	304
200K-300K	1
250K-300K	328
300K-400K	488
400K-500K	391
500K-600K	270
600-800K	363
800K-1M	233

Figure 11 – Final Data



	SALE TYPE	PROPERTY TYPE	PROPERTY TYPE 2	ADDRESS	CITY	CITY 2	STATE OR PROVINCE	ZIP OR POSTAL CODE	PRICE	BEDS	BATHS	LOCATION	SQUARE FEET
1	MLS Listing	Condo/Co-op	Condo/Co-op	2121 El Paseo St ...	Houston	Houston	TX	77054	114500	2	1	Cambridge Court...	804
2	MLS Listing	Condo/Co-op	Condo/Co-op	8704 Victorian Vil...	Houston	Houston	TX	77071	93000	2	1.5	Village Fondren ...	1184
3	MLS Listing	Condo/Co-op	Condo/Co-op	705 Main St #401	Houston	Houston	TX	77002	210000	1	1	St Germain Cond...	906
4	MLS Listing	Condo/Co-op	Condo/Co-op	1600 Post Oak Bl...	Houston	Houston	TX	77056	785000	2	2	Cosmopolitan Co...	1779
5	MLS Listing	Condo/Co-op	Condo/Co-op	7575 Kirby Dr #3...	Houston	Houston	TX	77030	158000	1	1	7575 Kirby	783
6	MLS Listing	Condo/Co-op	Condo/Co-op	7697 Cambridge St	Houston	Houston	TX	77054	189000	2	3	Medical Center T...	1368
7	MLS Listing	Condo/Co-op	Condo/Co-op	7575 Kirby Dr #1...	Houston	Houston	TX	77030	190000	1	1	7575 Kirby	1024
8	MLS Listing	Condo/Co-op	Condo/Co-op	2000 Bagby St #7...	Houston	Houston	TX	77002	188000	1	1	Rise Condo	586
9	MLS Listing	Condo/Co-op	Condo/Co-op	2400 Mccue Rd #...	Houston	Houston	TX	77056	210000	1	1	2400 Mccue Con...	801
10	MLS Listing	Condo/Co-op	Condo/Co-op	2001 Bering Unit ...	Houston	Houston	TX	77057	142500	1	1	2001 Bering	957
11	MLS Listing	Condo/Co-op	Condo/Co-op	9700 Leawood Bl...	Houston	Houston	TX	77099	75000	2	2	Leawood Condo ...	1068
12	MLS Listing	Condo/Co-op	Condo/Co-op	2277 S Kirkwood ...	Houston	Houston	TX	77077	165000	2	2	Oaks On Kirkwoo...	1155
13	MLS Listing	Condo/Co-op	Condo/Co-op	2121 El Paseo St ...	Houston	Houston	TX	77054	115000	2	1	Cambridge Court...	804
14	MLS Listing	Condo/Co-op	Condo/Co-op	2333 Bering Dr #...	Houston	Houston	TX	77057	122900	1	1	Park Regency Co...	837
15	MLS Listing	Condo/Co-op	Condo/Co-op	8433 Hearth Dr #2	Houston	Houston	TX	77054	75000	2	1.5	Hearthwood Con...	1040
16	MLS Listing	Condo/Co-op	Condo/Co-op	2816 S Bartell Dr ...	Houston	Houston	TX	77054	55000	1	1	Hearthwood 02 C...	700
17	MLS Listing	Condo/Co-op	Condo/Co-op	1880 White Oak ...	Houston	Houston	TX	77009	180000	2	2	White Oak Cond...	957
18	MLS Listing	Condo/Co-op	Condo/Co-op	4041 Law St #405	Houston	Houston	TX	77005	245000	2	2	Oaks/west Univer...	1042
19	MLS Listing	Condo/Co-op	Condo/Co-op	2100 Welch St U...	Houston	Houston	TX	77019	260000	2	1.5	Renaissance At Ri...	1244
20	MLS Listing	Condo/Co-op	Condo/Co-op	1500 Bay Area Bl...	Houston	Houston	TX	77058	107200	2	2	Baywind Condo S...	1035
21	MLS Listing	Condo/Co-op	Condo/Co-op	2255 Braeswood ...	Houston	Houston	TX	77030	184900	2	2	Braeswood Park ...	1096
22	MLS Listing	Condo/Co-op	Condo/Co-op	8100 Cambridge ...	Houston	Houston	TX	77054	115000	1	1	Cambridge Glen ...	1140
23	MLS Listing	Condo/Co-op	Condo/Co-op	10053 Westpark ...	Houston	Houston	TX	77042	75000	1	1	Idlewood Condo	676
24	MLS Listing	Condo/Co-op	Condo/Co-op	4521 San Felipe S...	Houston	Houston	TX	77027	879000	2	2.5	Arabella Condo A...	1579
25	MLS Listing	Condo/Co-op	Condo/Co-op	1515 Sandy Sprin...	Houston	Houston	TX	77042	109900	1	1.5	Lynbrook Manor ...	1056
26	MLS Listing	Condo/Co-op	Condo/Co-op	2800 Jeanetta St ...	Houston	Houston	TX	77063	124900	2	2	One Orleans Plac...	1043
27	MLS Listing	Condo/Co-op	Condo/Co-op	6320 Crab Orchar...	Houston	Houston	TX	77057	234900	2	2	Kerry Glen Cond...	1513
28	MLS Listing	Condo/Co-op	Condo/Co-op	2016 Main St #22...	Houston	Houston	TX	77002	175707	1	1	Main Condo 03 A...	697
29	MLS Listing	Condo/Co-op	Condo/Co-op	7900 Westheimer...	Houston	Houston	TX	77063	108500	2	2	Dillston House Co...	985
30	MLS Listing	Condo/Co-op	Condo/Co-op	1515 Sandy Sprin...	Houston	Houston	TX	77042	105000	2	2.5	Lynbrook Manor ...	1225
31	MLS Listing	Condo/Co-op	Condo/Co-op	2710 Hullsmith D...	Houston	Houston	TX	77063	245000	2	3	Contemporary W...	1501
32	MLS Listing	Condo/Co-op	Condo/Co-op	14911 Wunderlic...	Houston	Houston	TX	77069	124900	2	2	Bridgewater Land...	1068
33	MLS Listing	Condo/Co-op	Condo/Co-op	1409 Post Oak Bl...	Houston	Houston	TX	77056	1300000	3	3.5	Astoria Condos	3010
34	MLS Listing	Condo/Co-op	Condo/Co-op	2121 Hepburn St ...	Houston	Houston	TX	77054	118000	2	2	Montreal Place C...	1046
35	MLS Listing	Condo/Co-op	Condo/Co-op	9513 Pagewood Ln	Houston	Houston	TX	77063	106990	2	2	Tanglewilde Sout...	1224
36	MLS Listing	Condo/Co-op	Condo/Co-op	9707 Richmond ...	Houston	Houston	TX	77042	95000	1	1	Doma Chase Con...	836
37	MLS Listing	Condo/Co-op	Condo/Co-op	2121 El Paseo St ...	Houston	Houston	TX	77054	142000	2	1	Cambridge Court...	804
38	MLS Listing	Condo/Co-op	Condo/Co-op	351 N Post Oak L...	Houston	Houston	TX	77024	129000	1	1	Post Oak Lane	894