Analysis of Zillow Data from Austin, Texas in November 2022

Will Curkan

2022-12-16

Introduction

Austin, a hilly city due south in the middle of Texas, is where all the young, hip Texans and out-of-staters are converging to live. Other than its famous 6^{th} street nightlife, the "weird" city is home to many "corporate and regional headquarters" of companies like Dell, IBM, Facebook, VISA and many more [1] - these are some great reasons to "kick up the roots" and mosey on over to Austin! Suppose you think that you are *just* young and hip enough to move out to Austin, you might want some insights on how the housing market is.

Objective Statement

The objective of this project is to explore the dataset, make inferences about true statistic parameters in Austin, Texas, and to make predictions on housing prices and categories like "which broker is listing the property".

Data Acquisition and Explanation

The user Alex Huggler on Kaggle.com provides Zillow Housing data for November 2022 [2]. The data has 800 unique observations and 20 features: one observation per unique housing unit and each feature as some trait of the house like its zip code or number of bedrooms.

Feature Descriptions

The following features will be used in analysis of Austin housing prices:

Feature	Description
price	Price of housing
addressCity	City name
addressZip	Zip code
beds	Number of beds
baths	Number of baths
area	Sq. footage
latitude	Latitude
longitude	Longitude
variableData	String of description from Zillow
sgapt	For Sale by broker OR new construction (categ.)
zestimate	Zillow price estimate
brokerName	Broker (seller) name

Features from the dataset have been removed for this analysis. The feature `unformattedPrice` is removed because it is the same as `price`. `address` and `addressStreet` are removed as they will not be useful in analysis or prediction because they're all unique and non-numeric. `addressState` only had the value "Texas". `isZillowOwned` is only FALSE as Zillow does not own any of the properties. `badgeInfo` is undefined and is NULL. `pgapt` is all "For Sale". `info3string` is a link to a broker logo pictures.

EDA

Checking Cities

We did not need the `addressState` column because it only had one value: Texas. We might assume that `addressCity` is only Austin and remove the feature, but is there another city in the dataset?

City	Listings
Austin	786
Del Valle	8
Pflugerville	3
Lakeway	1
Manor	1
Terlingua	1

There are a few other cities represented in this dataset, so we will keep this column for now.

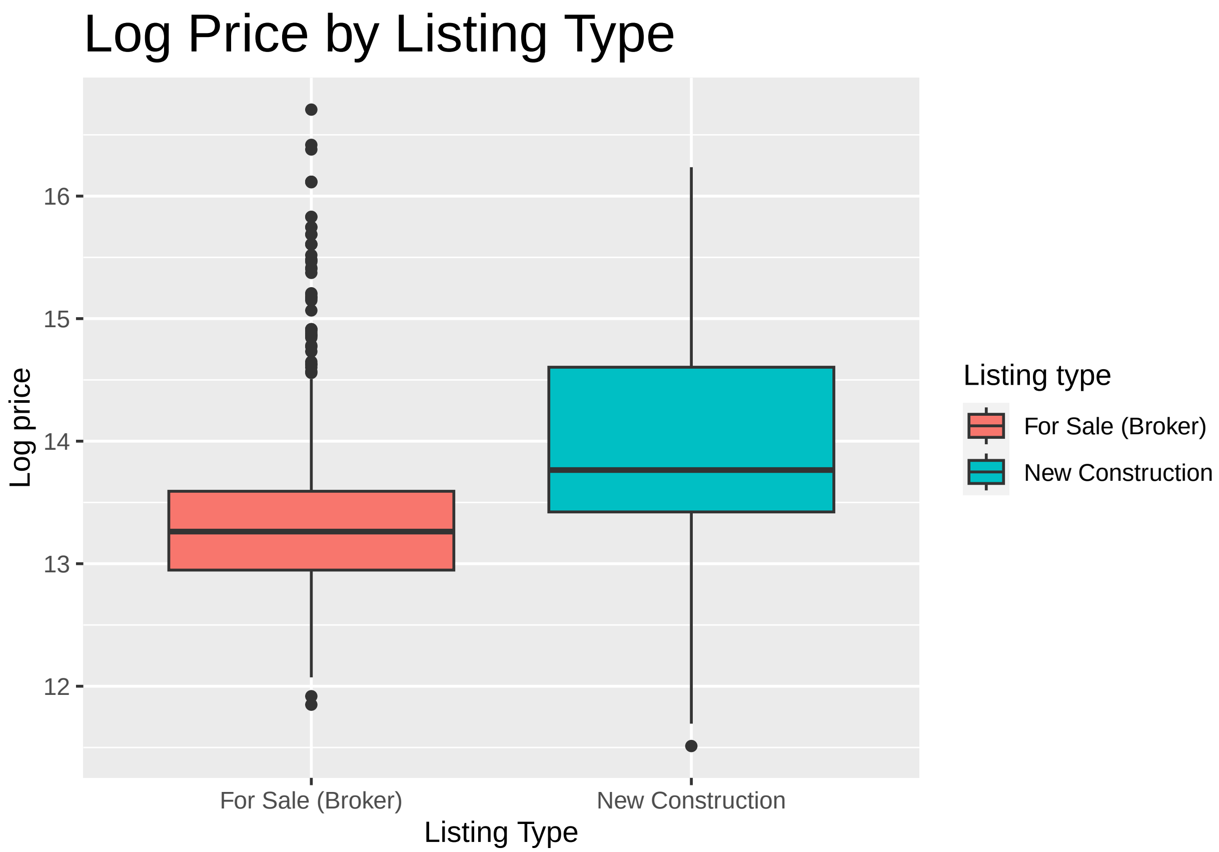
Brand-New or Old house

It is also interesting to know which houses are being sold and under construction, or are for sale by a broker.

Listing type	Listings
For Sale (Broker)	755
New Construction	45

Most of the houses for sale have already been built. Lets see if there is a price difference between the types of listings.

Note: Log price is used instead of price to better visualize the similarity in scale of prices.



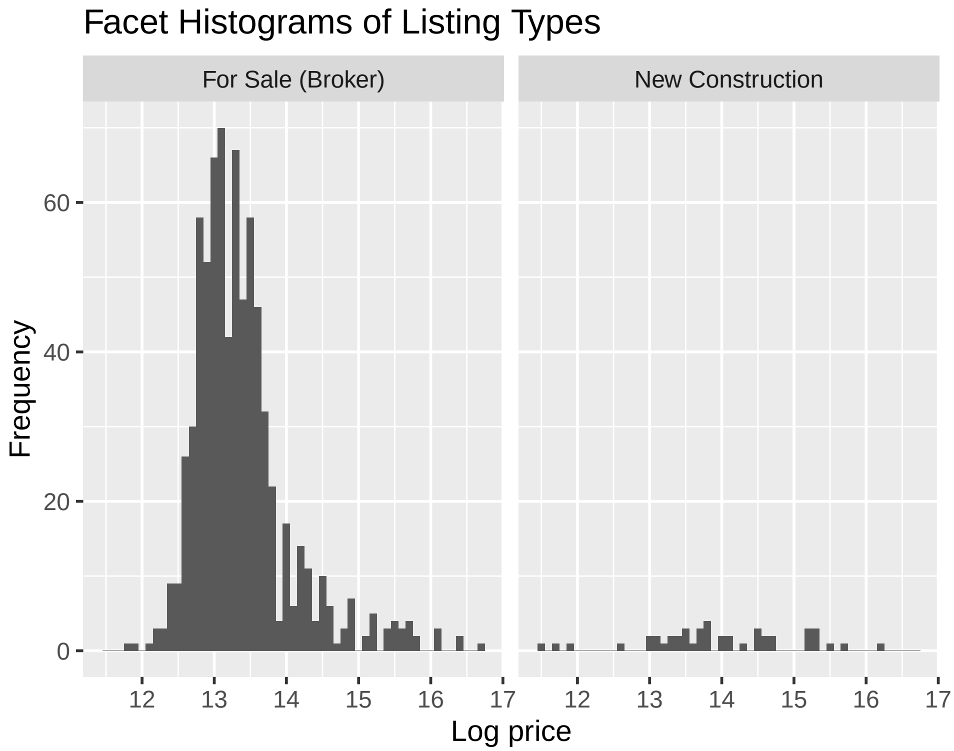
Looks like buying a new house tends to cost more than one already for sale by a broker.

So, let's look at the mean housing price of both listing types.

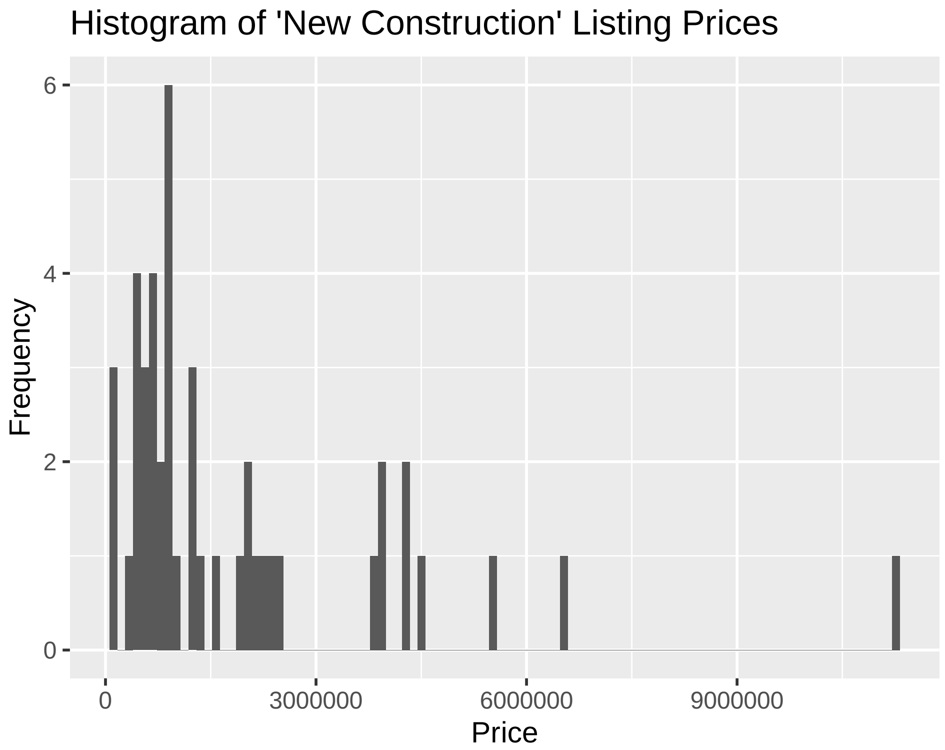
Listing type	Mean price
For Sale (Broker)	915,297.9
New Construction	1,882,131.2

Looking at the mean price of the listing types is disheartening, not only do you have to be hip to live in Austin, but you have to be rich too? The mean prices of the two listing types are quite high, and we see from the boxplot that the data is skewed, so let's employ some resampling techniques to get an estimate on the true mean house prices.

It might be worth taking a closer look at the normality of the data.

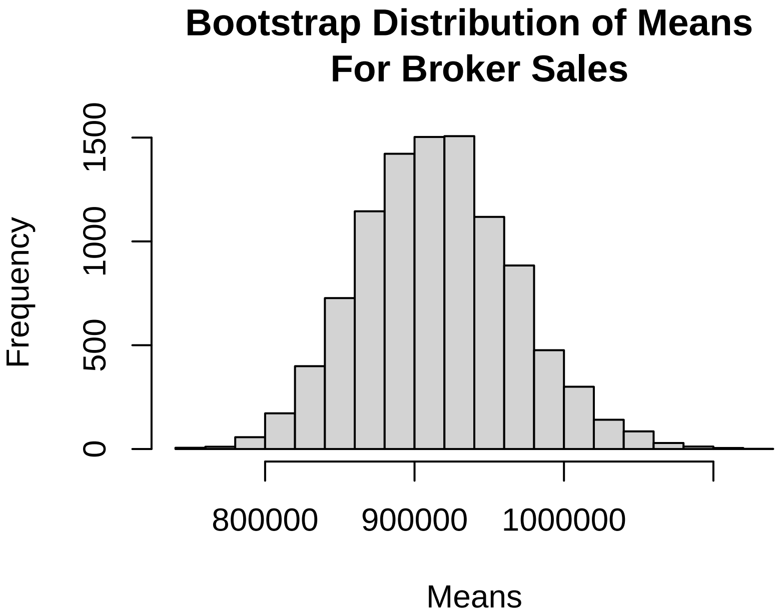


We can already observe the skewedness of the Broker sale data from the boxplot, but it is hard to tell the shape of the “New Construction” data.

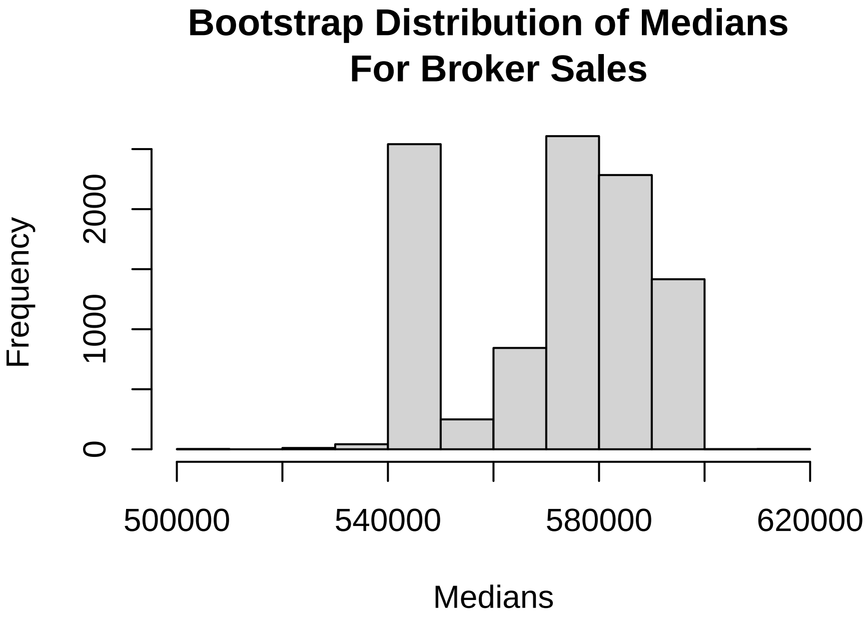


So, there are quite a few houses over the million dollar price range... And the most expensive house is $11,250,800. These are probably for Elon Musk and his pals, though.

Let's get a range estimate for the true mean house prices of both listing types by looking at the distribution of the bootstrap means.

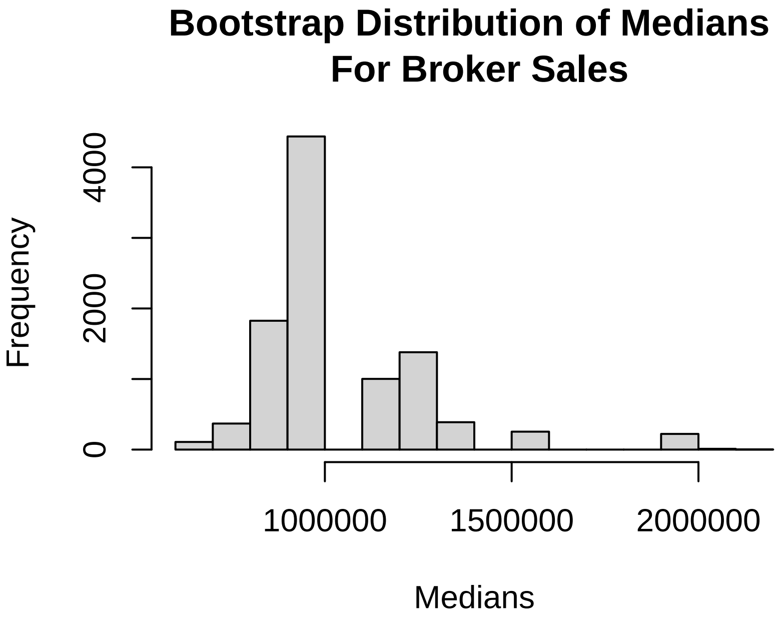


We are 95% confident in saying that the true mean housing price in Austin is between (820715.3 1021421.0) and our bias percentage shows the sample mean may well represent the true mean, with the bias being only .6%. But we actually want people to move to Austin and the mean isn't the best descriptor in this case, so let's check out the bootstrap distribution of the median and see if we get a more reasonable price.



We are 95% confident that (549000 599000) is one of the infinitely many 95% CIs that contain the true median house price. Our bias percentage is kind of high, at 10.89% biased. Anyway, around 500k - 600k for a house sounds a lot better than an average of 900k.

How does this compare to newly constructed houses? We saw that the average price for newly constructed housing units was approximately double that of ones for sale by a broker. Since the median looks like a better estimator, let's look at the median for newly constructed houses.



The median is highly skewed, the interval estimate from the 95% CI has an enormous range, and the bias percentage is approximately 50% (.4786609), so it is not wise to trust population parameters for new houses coming from this sample.

Let's perform one more test on median housing price. The website realtor.com claims that, in November 2022, the median housing price was 600k USD [3]. If the price is 600k or less, that is great for buyers. If not, then their claim is not accurate. We will use both listing types for the test. We will test the hypotheses:

$H_0 : \tilde{\mu} <= 600,000$ versus $H_A : \tilde{\mu} > 600,000$

The p-value, being the probability that the statistic occurs randomly, is approximately 50.32%. Given the data, we fail to reject the null hypothesis at any small level of significance α and say that *there is not statistically significant evidence to contradict the claim that the median house price is more than 600k.* Unfortunately, the bias percentage is quite high: around 27%. But we're fairly certain we could find some options under 600k. Austin seems like "the spot", though, so we may be willing to dish out some extra Quiche (cash) for a "spot at the spot".

Top Brokers

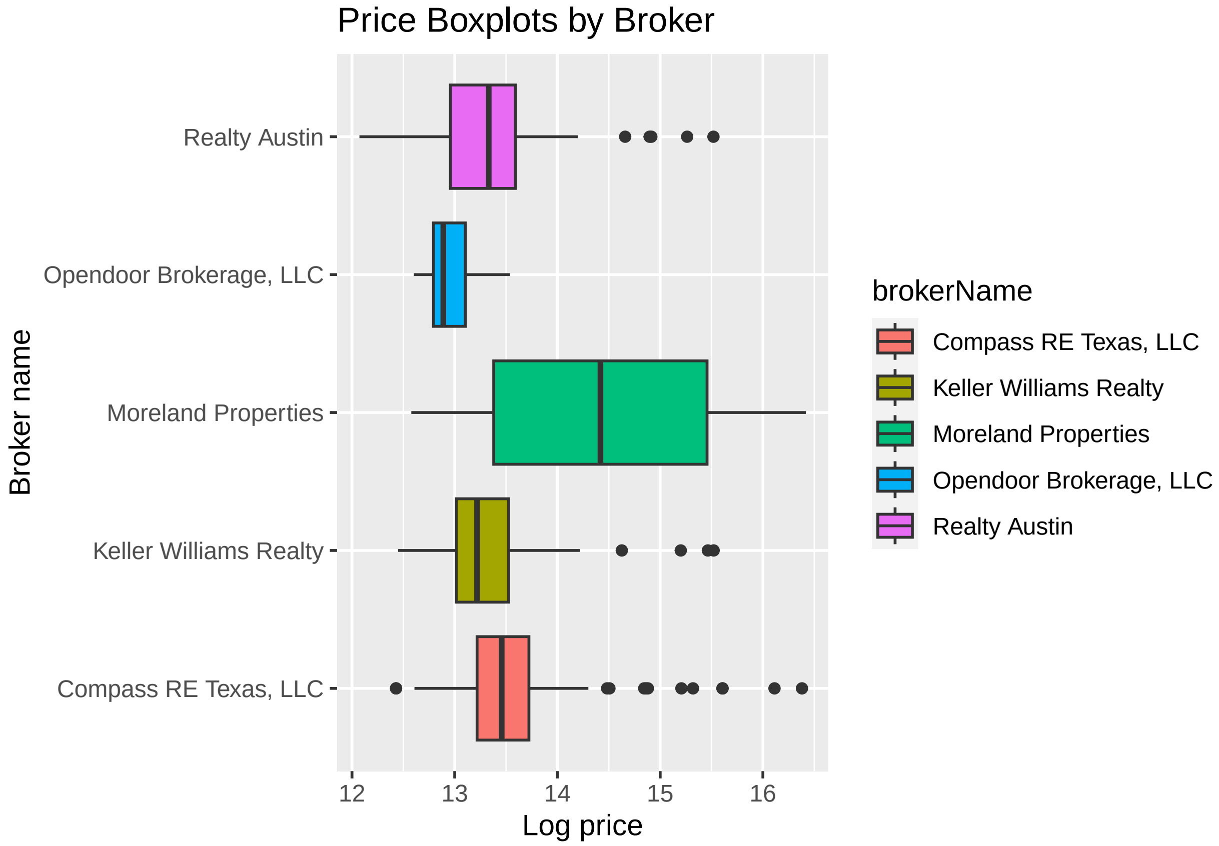
Of course, life isn't all butterflies and candy: we have to make Quiche Lorraine, buy property, maintain the property... I digress.

The bias percentages were quite high for the types of listings, so it may not truly matter if we choose newly constructed houses or older ones; let's just look at choosing a broker.

We think maybe we should go with the brands selling the most properties, as the buying process may be easier and the price may be lower. The Brokers with the most listings are:

Broker	Listings
Realty Austin	73
Keller Williams Realty	67
Compass RE Texas, LLC	60
Opendoor Brokerage, LLC	32
Moreland Properties	29

Other than Keller Williams Realty, the other brokers are not huge names if you are not into the housing market. Let's explore the data relative to the top 5 brokers.



It looks like Moreland Properties has the highest variation in house prices. Moreland also has the fewest listings, likely because they're so high priced. If you are moving to the greater Austin area and your friend recommends using Moreland Properties, they probably think you are rich!

Keller Williams, Austin Realty, and Compass RE look like they are doing similar business, though Compass has some huge outliers.

The median house prices by broker:

Broker	Median Price
Realty Austin	616000
Keller Williams Realty	550000
Compass RE Texas, LLC	699000
Opendoor Brokerage, LLC	396000
Moreland Properties	1830000

The median for all brokers, besides Moreland Properties, seem reasonable for Austin, Texas, given the claim from realty.com holds true for November 2022.

Getting the most bang-for-our-buck is a good idea as settling down and having a family is typical human behavior. Let's look to do the following:

1. Find the cost per square footage

- Create a new feature that is the ratio of area (square footage) to the price; USD cost per square foot; usd_per_sqfoot
- $usd\_per\_sqfoot = \frac{area}{price}$

2. Find the ratio of the cost per square foot (usd_per_sqfoot) to the full housing price

- Create a new feature that is the $cost\_ratio = \frac{usd\_per\_sqfoot}{price}$

We need to find the USD price per square foot, but we cannot stop there because a low USD price does not mean we are getting a huge space; the house could be horribly small and that's why it is so cheap. We should aim for a low usd_per_sqfoot to price ratio; the more area we get per bigger-price -tagged house will have a cheaper USD to square foot, and the lower ratio of the USD per square foot to the total housing price will give us a lower ratio meaning we minimize cost per square foot and maximize the area.

A house with a low `cost_ratio` is preferable.

```
##       price    cost_ratio  area beds baths usd_per_sqfoot addressCity
## 1  18000000 0.00007130125 14025    3     8      1283.4225      Austin
## 2   5200000 0.00007670476 13037    6    10       398.8648      Austin
## 3   6499000 0.00009401147 10637    6     7       610.9805      Austin
## 4   3900000 0.00009703086 10306    7     9       378.4203      Austin
## 5  13500000 0.00011697275  8549    5     7      1579.1321      Austin
## 6   9990000 0.00011966017  8357    5     8      1195.4050      Austin
## 7  11250800 0.00012583365  7947    5     8      1415.7292      Austin
## 8   7500000 0.00012894907  7755    5     6       967.1180      Austin
## 9   7499000 0.00012919897  7740    6     8       968.8630      Austin
## 10  2995000 0.00014285714  7000    6     6       427.8571      Austin
```

It turns out the most expensive houses are the best deal on the ratio of cost per square foot to the total cost, so we should set a limit that we only want to look at houses under, say 750,000$

```
##                                      address
## 1      12617 Black Hills Dr, Austin, TX 78748
## 2   Plan 3475 Modeled Plan, McKinney Crossing
## 3        11206 Kingsgate Dr, Austin, TX 78748
## 4         6705 Ondantra Bnd, Austin, TX 78744
## 5     7125 Wandering Oak Rd, Austin, TX 78749
## 6     6849 Thistle Hill Way, Austin, TX 78754
## 7  6917 William Wallace Way, Austin, TX 78754
## 8       8017 Bottlebrush Dr, Austin, TX 78750
## 9             15014 Iowa St, Austin, TX 78734
## 10            15901 Arla Cv, Austin, TX 78717
```

Now we're talking. We see some big, median-priced houses. Of course it may be smarter to go cheaper, but this works for now.

End note: The `area` feature should be filled intelligently because the analysis isn't as good as it could be.

References:

[1] https://www.austinchamber.com/economic-development/key-industries/corporate-headquarters

[2] https://www.kaggle.com/datasets/alexhuggler/austin-zillow-houses-for-sale

[3] https://www.realtor.com/realestateandhomes-search/Austin_TX/overview