



USA Real Estate Dataset

LARRY SERRATOS

Objetivos

- Price Prediction

I like data science and im interested in learning more about real state, so i mix up this topics in my proyect to keep track of both at the same time.

I want to know if we can predict housing prices based on the features of a house.

So the objective is create a predictor model for the Price of a property in the US based on basic property information.

In the future I would like to mix this model with others made up of other datasets to include other characteristics of the properties and make better prediction for prices in the present and possibilities of investment (future prices).

Exploratory Data Analysis (EDA)

- Understanding the dataset
- Data structure
- Previsualization
- Outlier analysis
- Handle of missing data
- Correlation
- Relation of features and price

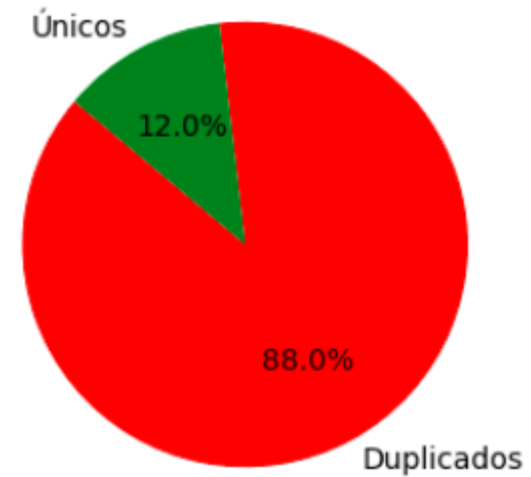
This dataset contains Real Estate listings in the US broken by State and zip code.

	status	bed	bath	acre_lot	city	state	zip_code	house_size	prev_sold_date	price
0	for_sale	3	2	0.12	Adjuntas	Puerto Rico	601	920	nan	105000
1	for_sale	4	2	0.08	Adjuntas	Puerto Rico	601	1527	nan	80000
2	for_sale	2	1	0.15	Juana Diaz	Puerto Rico	795	748	nan	67000
3	for_sale	4	2	0.1	Ponce	Puerto Rico	731	1800	nan	145000
4	for_sale	6	2	0.05	Mayaguez	Puerto Rico	680	nan	nan	65000

- status (Housing status - a. ready for sale or b. ready to build)
- bed (# of beds)
- bath (# of bathrooms)
- acre_lot (Property / Land size in acres)
- city (city name)
- state (state name)
- zip_code (postal code of the area)
- house_size (house area/size/living space in square feet)
- prev_sold_date (Previously sold date)
- price (Housing price, it is either the current listing price or recently sold price if the house is sold recently)

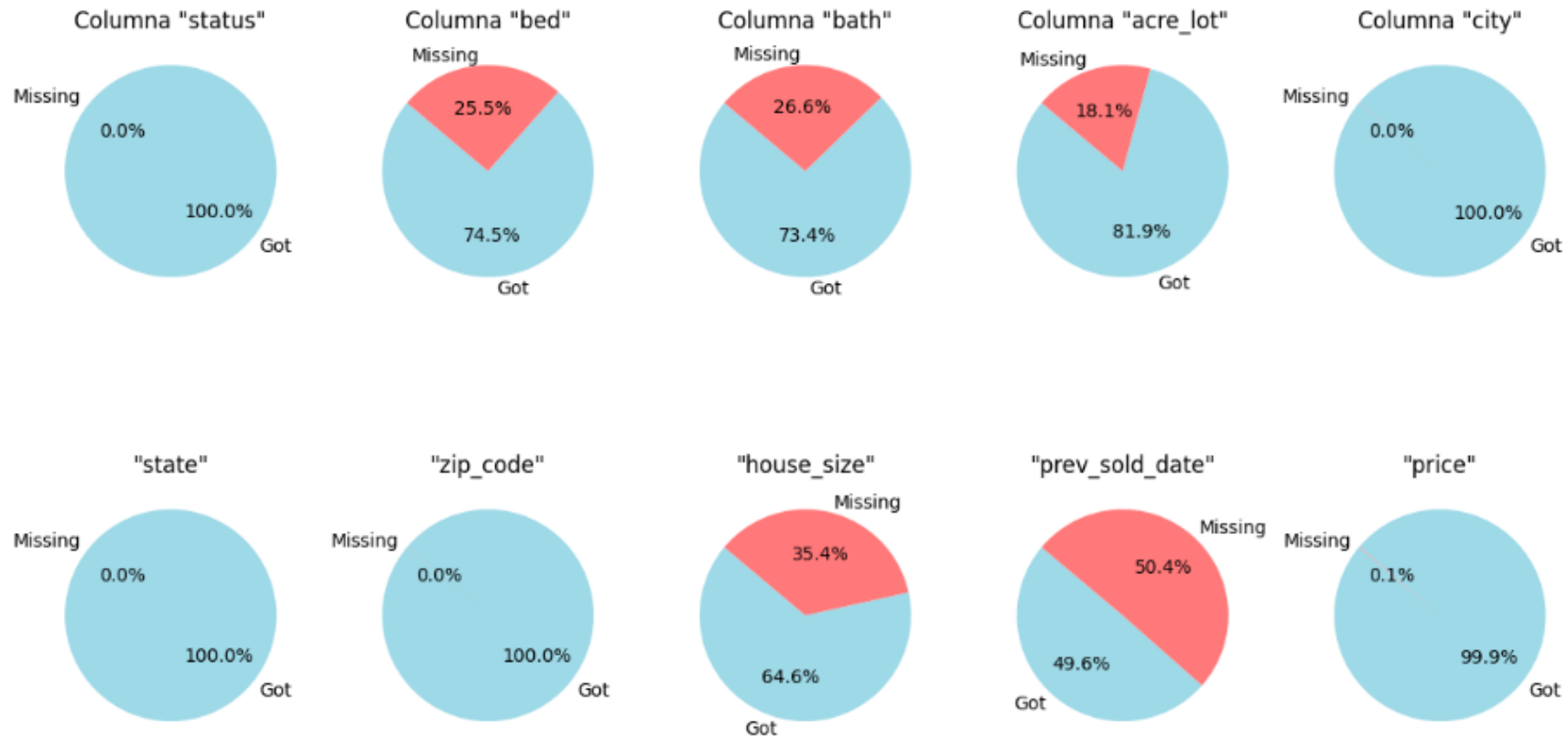
First filters

There are a lot of data that is duplicated.



There are still 424121 rows after dropping duplicates

Missing Values



Need to imputation to handle the missing values or drop them.

Previsualization of the outliers

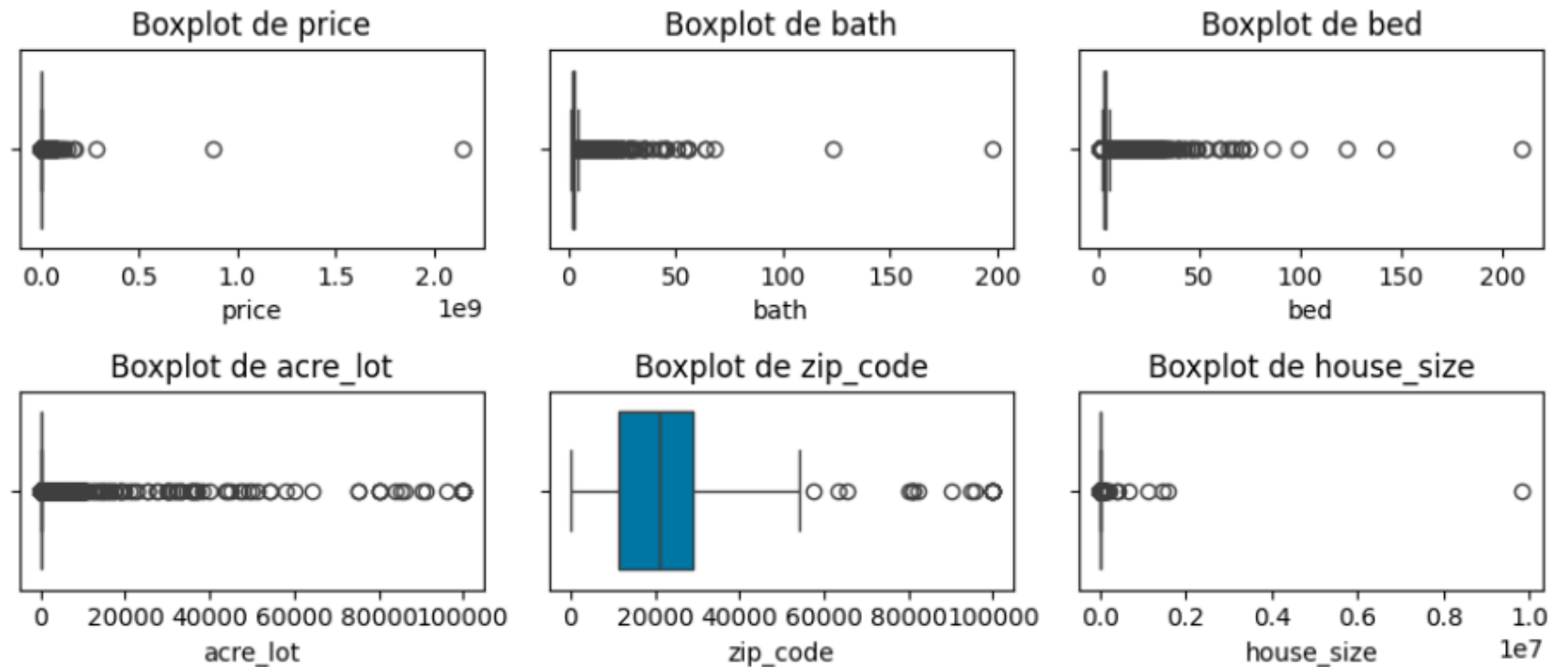
We can see some outliers in the price, probably related to the house size.

Drop prices up to 0.5×10^9

Drop rows with up to 100 bath

Drop up to 100 bed

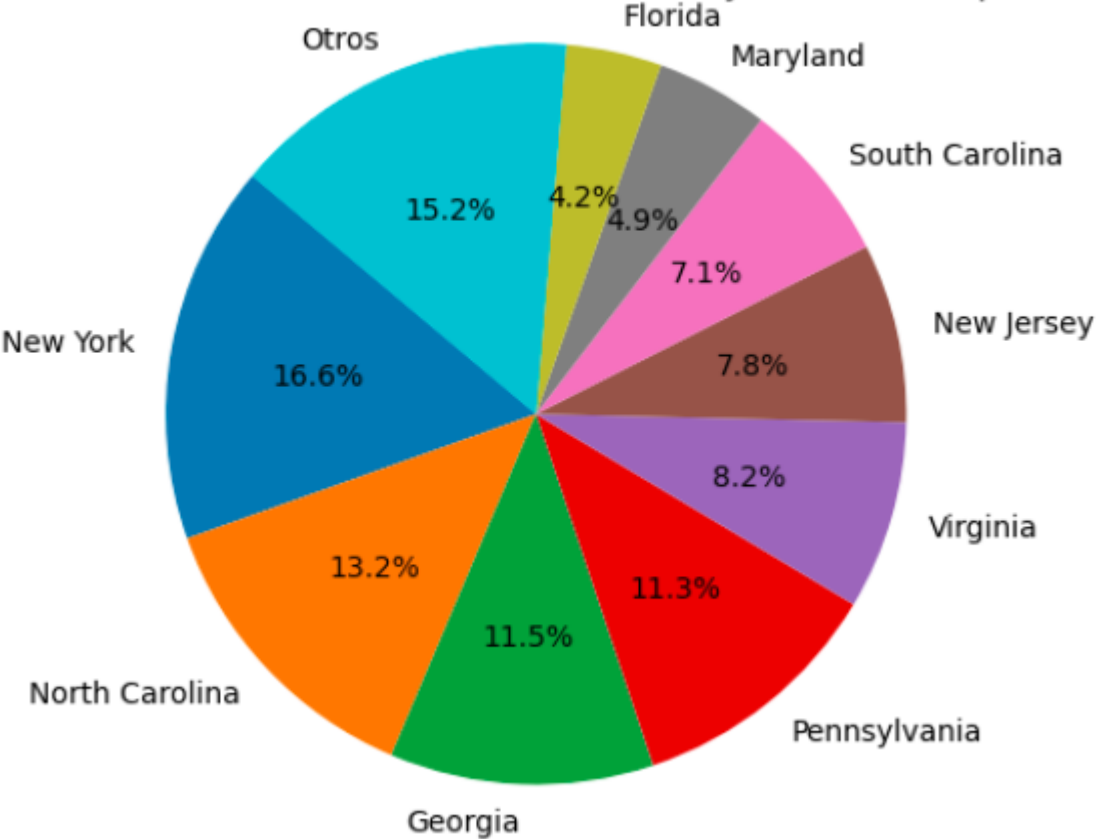
Drop up to 0.2 house size



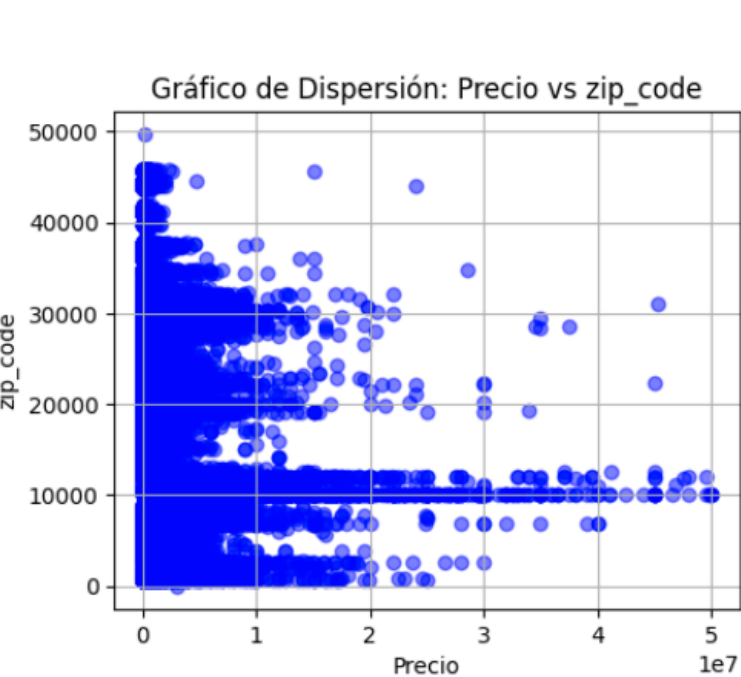
Previsualization of the different states

1.	Puerto Rico	12.	Virginia	23.	Maryland
2.	Virgin Islands	13.	Wyoming	24.	Missouri
3.	Massachusetts	14.	Maine	25.	District of Columbia
4.	Connecticut	15.	Georgia	26.	Wisconsin
5.	New Hampshire	16.	Pennsylvania	27.	North Carolina
6.	Vermont	17.	West Virginia	28.	Kentucky
7.	New Jersey	18.	Delaware	29.	Michigan
8.	New York	19.	Louisiana	30.	Mississippi
9.	South Carolina	20.	Ohio	31.	Florida
10.	Tennessee	21.	California	32.	Alabama
11.	Rhode Island	22.	Colorado	33.	New Brunswick

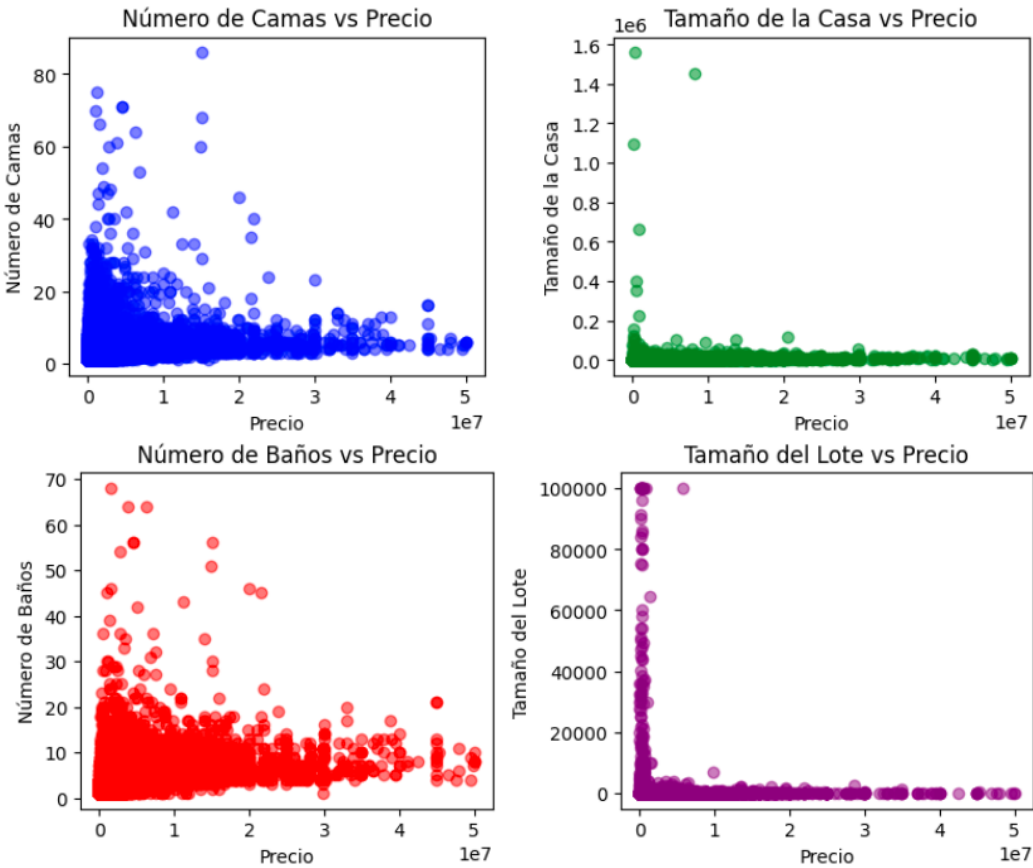
Distribución de los 9 Estados Más Comunes y Otros, sin duplicados



Previsualization of the average prices in the

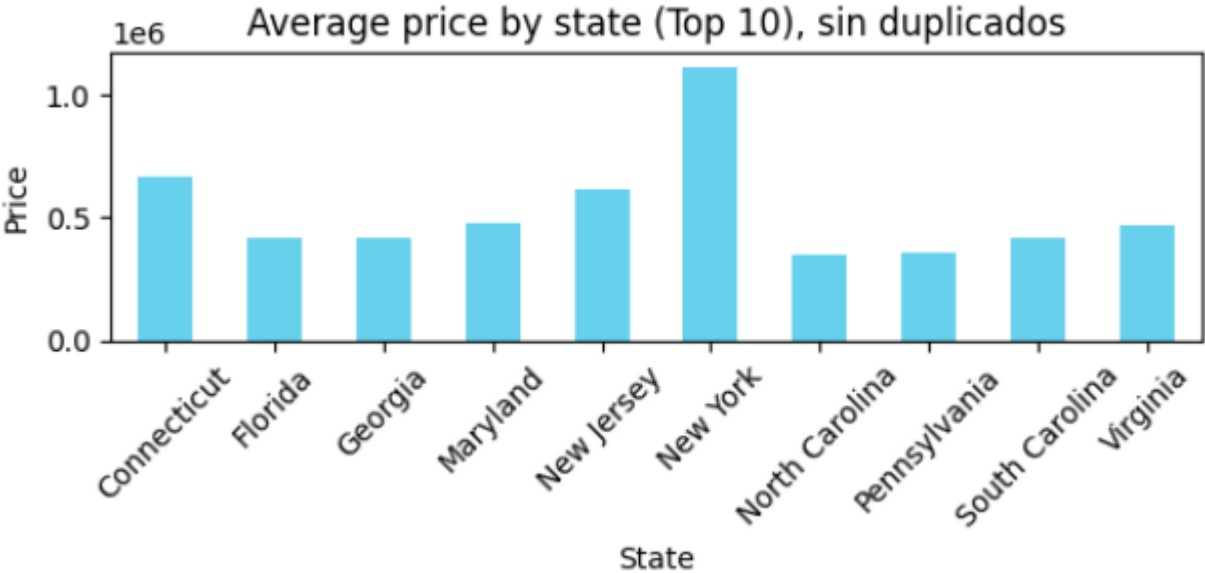


The price increase in zones of zip codes.



Previsualization of the average prices in the

The average prices could be very different according to the city.

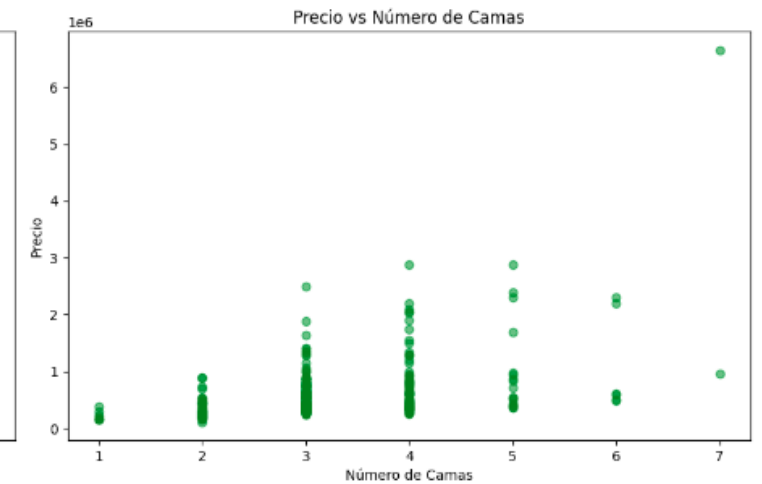
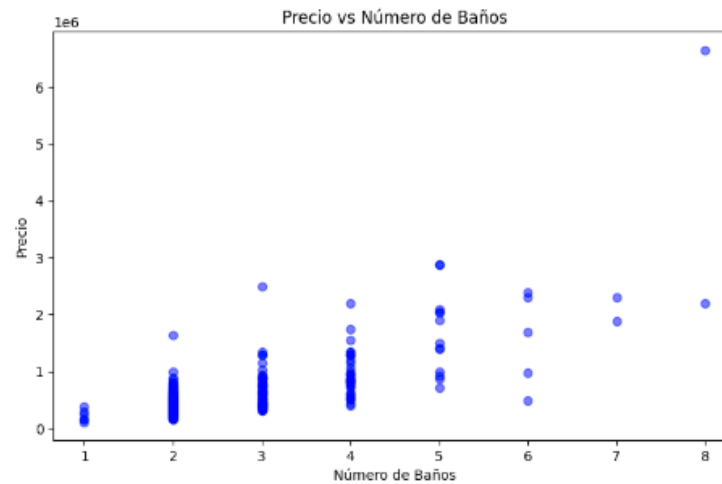


Los 5 códigos postales más caros con sus respectivas ciudades, estados y precios promedio:

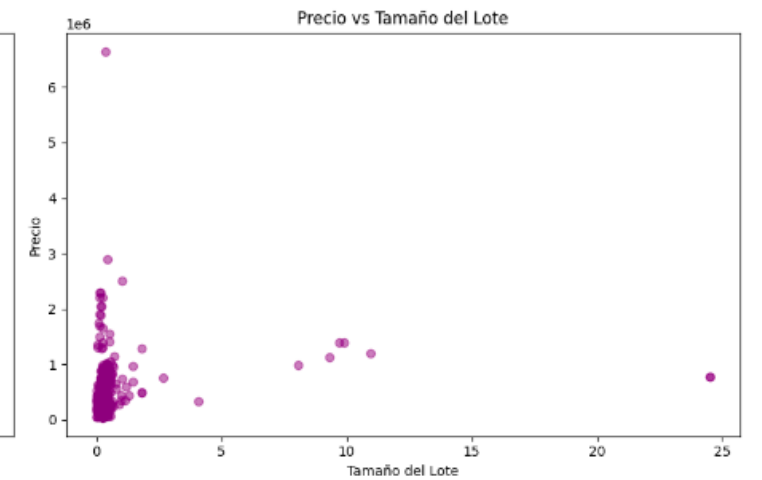
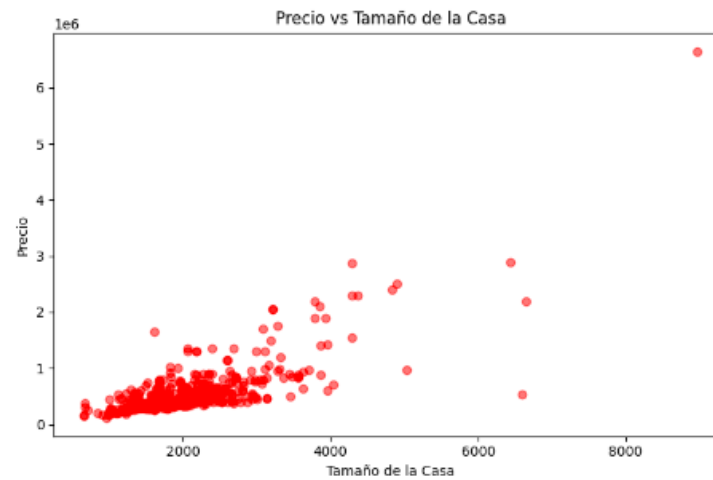
	zip_code	price	city	state
4105	10985.0	3.000000e+07	Thompson Ridge	New York
791	2672.0	1.810000e+07	Barnstable	Massachusetts
2453	6753.0	1.333000e+07	Cornwall	Connecticut
4821	11962.0	1.291905e+07	Sagaponack	New York
717	2543.0	1.230000e+07	Woods Hole	Massachusetts

Analysis of features respect to a single zip_code

We can notice the linear growing of the price in proportion to the features.



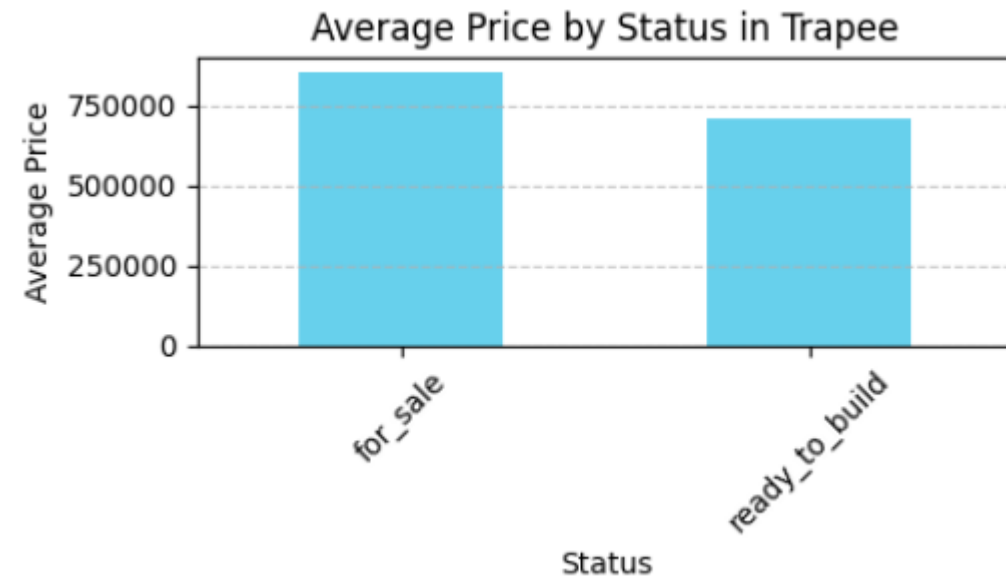
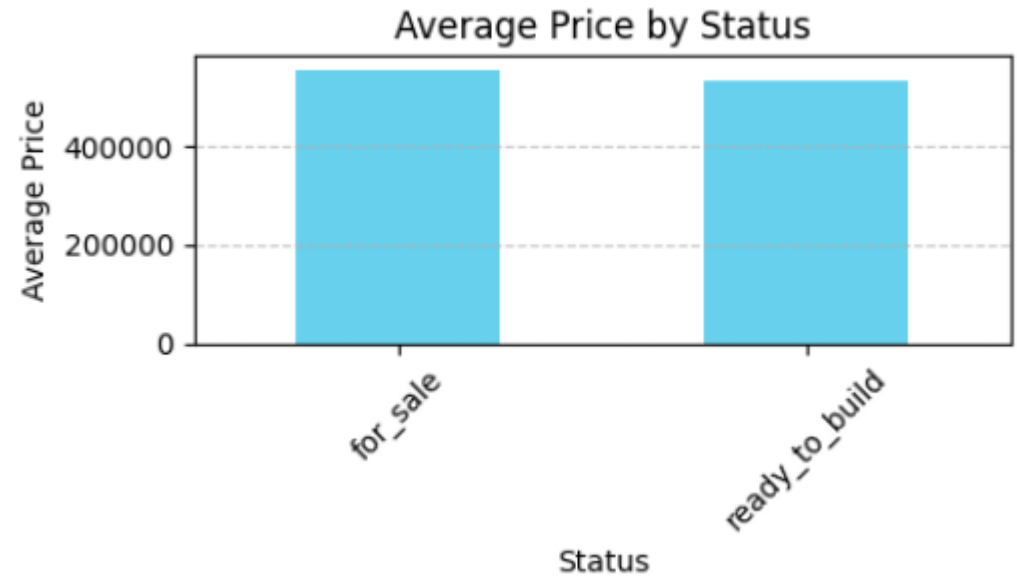
The acre lot is not



Analysis of features respect to a single zip_code

We can notice the linear growing of the price in proportion to the features.

The acre lot is not



Dataset

- The data set have a lot of duplicated values
- The data set have a lot of missing values that need imputation or dropping
- Check the outliers

Statements

- The zip_code is the best predictor, but not in a linear way, more like in a neighborhood way.
- Bath, bed and size of house are secondary good predictors. Not as good as acre_lot

Other possible insights

- Does the duration of ownership impact the price of the house?

La mejor cita que refleja su visión... "Es un pequeño paso para el hombre, pero un gran salto para la humanidad".

-NEIL ARMSTRONG