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 objetive 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 posibilities 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.

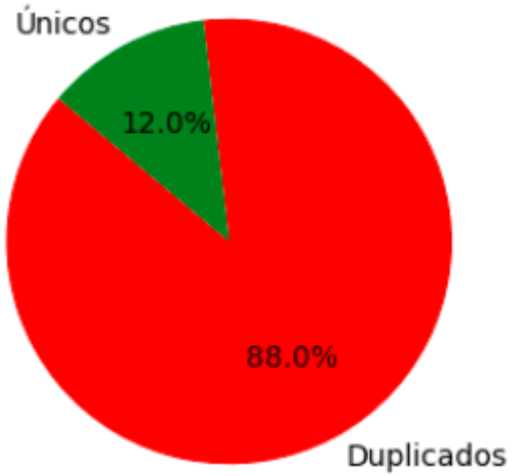
	status	bed	bath	acre_lot	city	state	zip_code	house_size	prev_sold_date	price
0	for_sale	3	2	0.12	Adjuntas	Puerto Rico	601	920	nan	105000
1	for_sale	4	2	0.08	Adjuntas	Puerto Rico	601	1527	nan	80000
2	for_sale	2	1	0.15	Juana Diaz	Puerto Rico	795	748	nan	67000
3	for_sale	4	2	0.1	Ponce	Puerto Rico	731	1800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

Data Structure

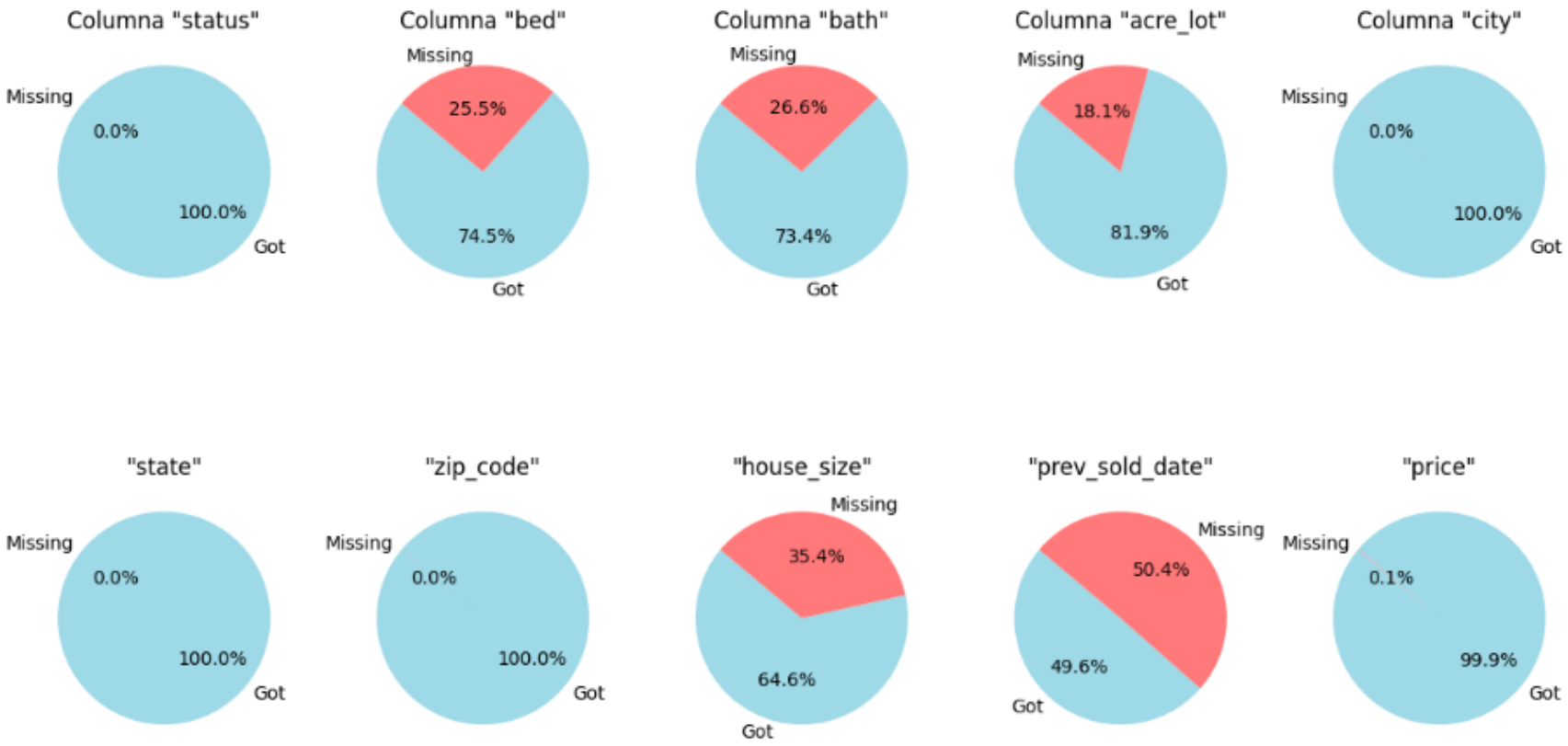
First filters

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

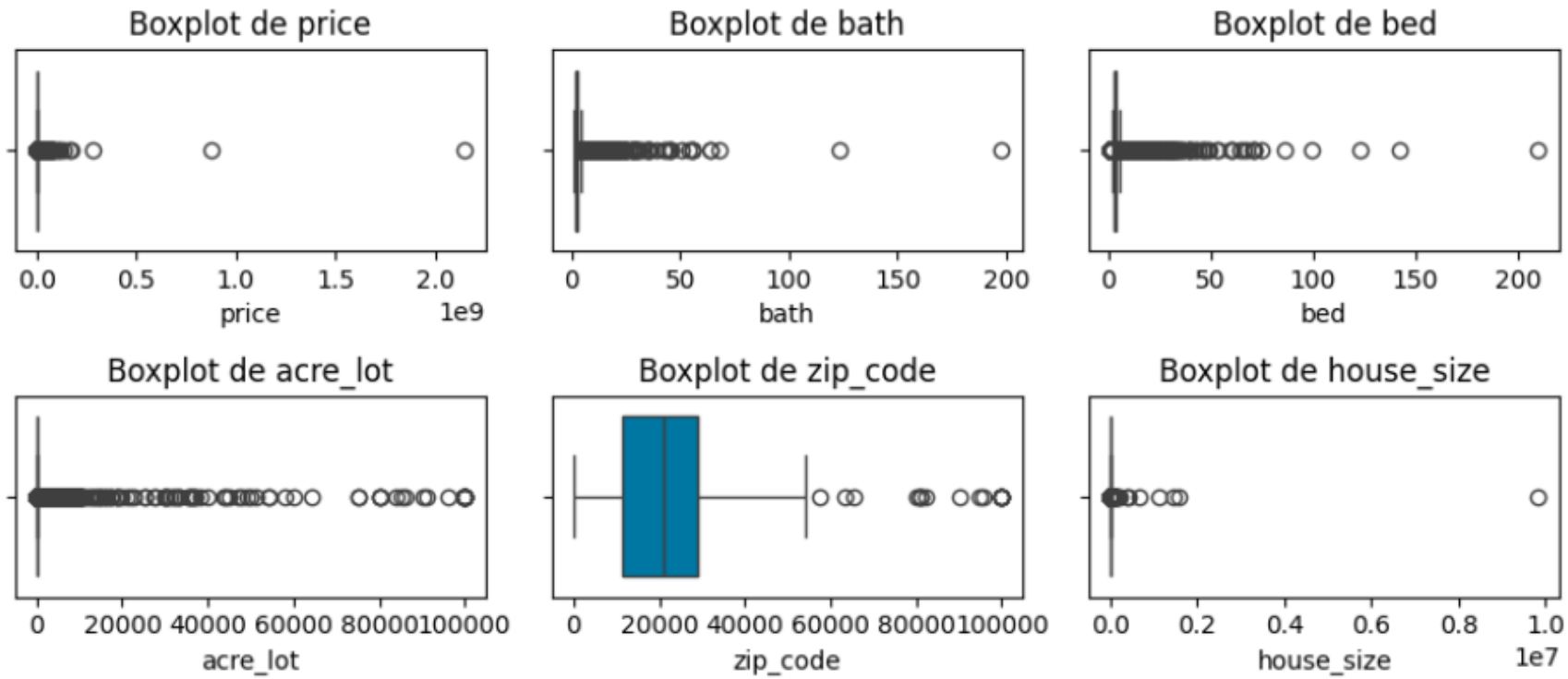
We can see some outliers in the price, probably related to the house size.

Drop prices up to 0.5*10^9

Drop rows with up to 100 bath

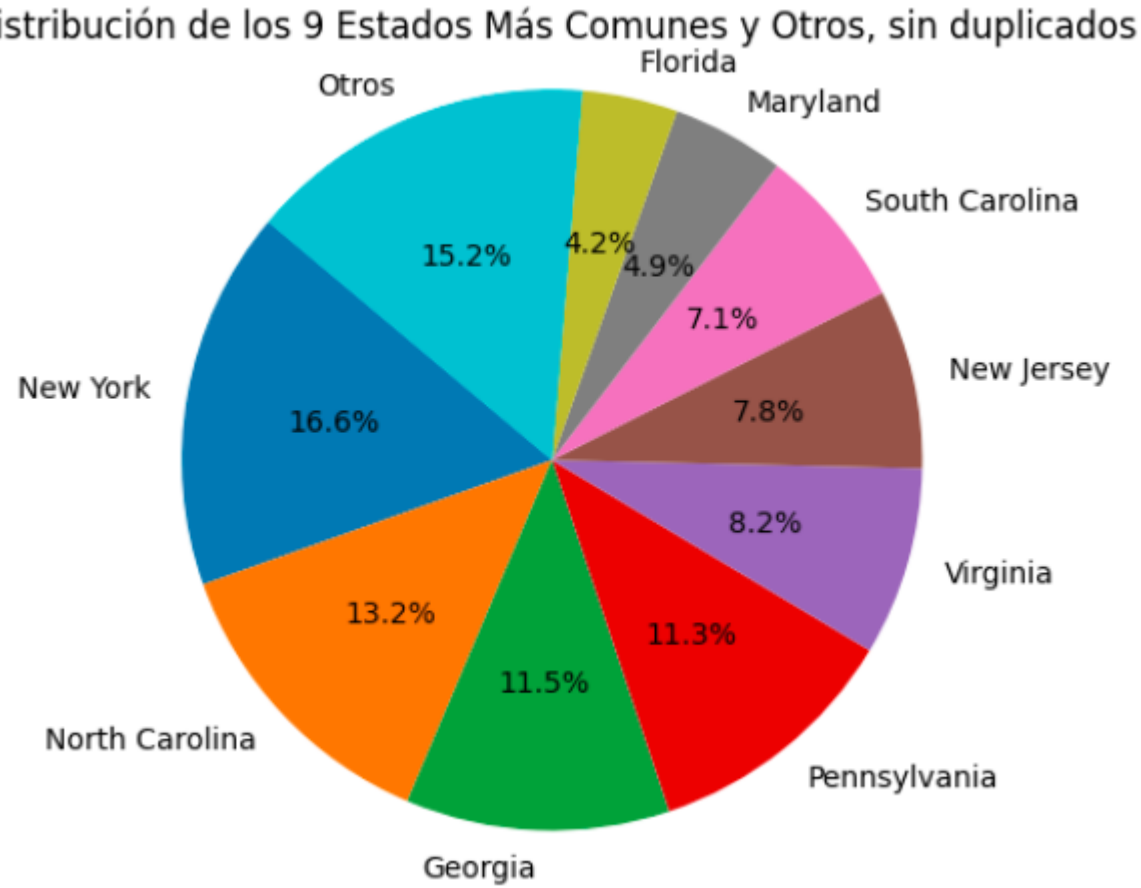
Drop up to 100 bed

Drop up to 0.2 house size

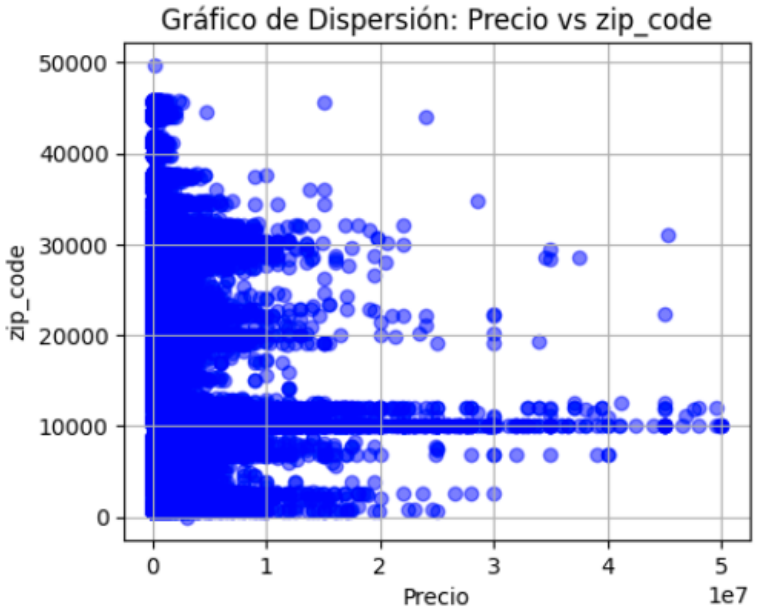


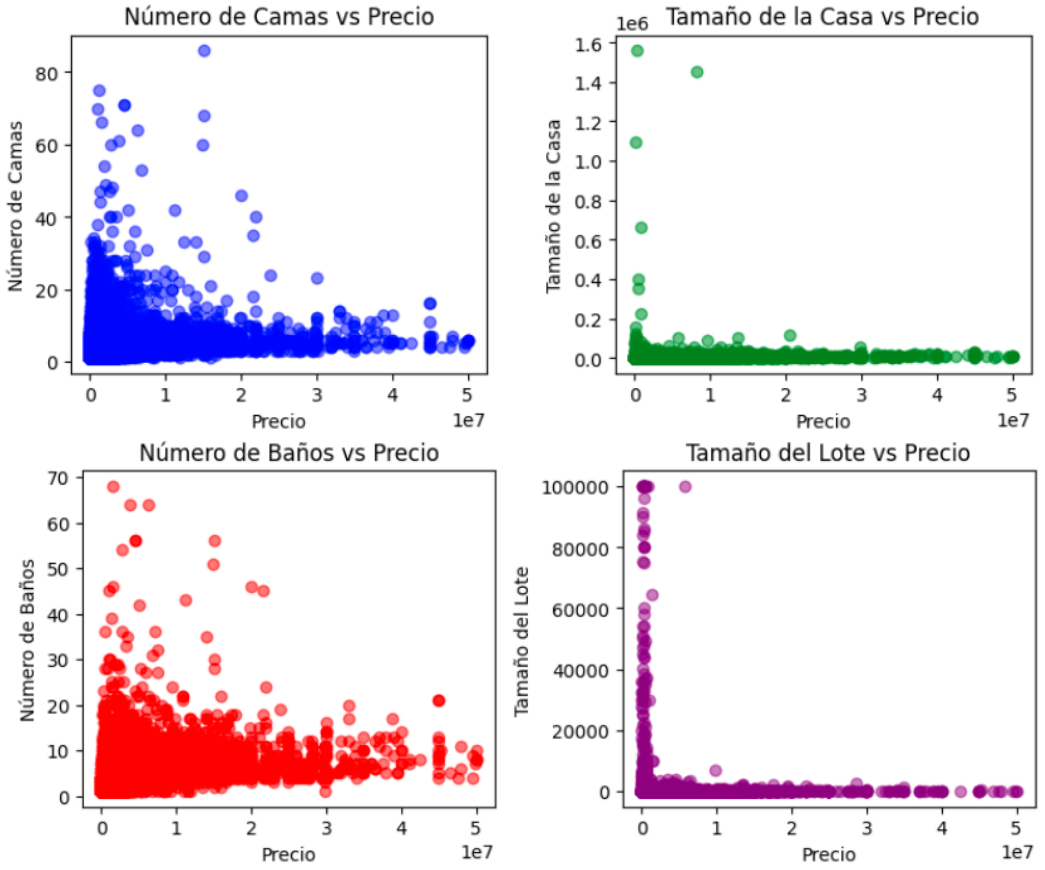
Previsualization of the different states

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



Previsualization of the average prices in the

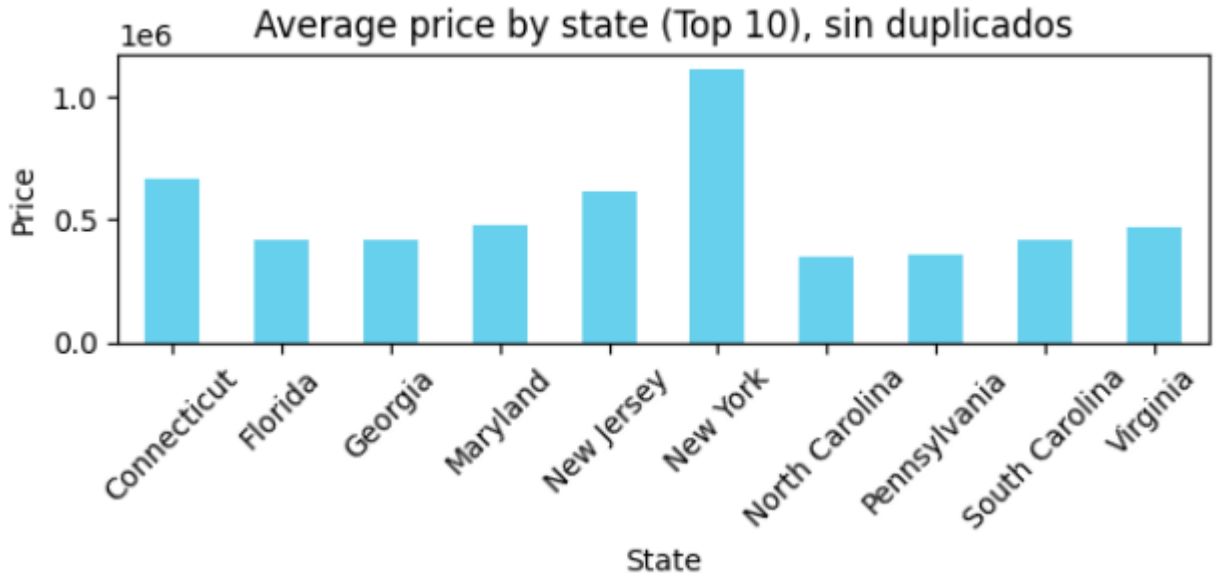




The price increase in zones of zip codes.

Previsualization of the average prices in the

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



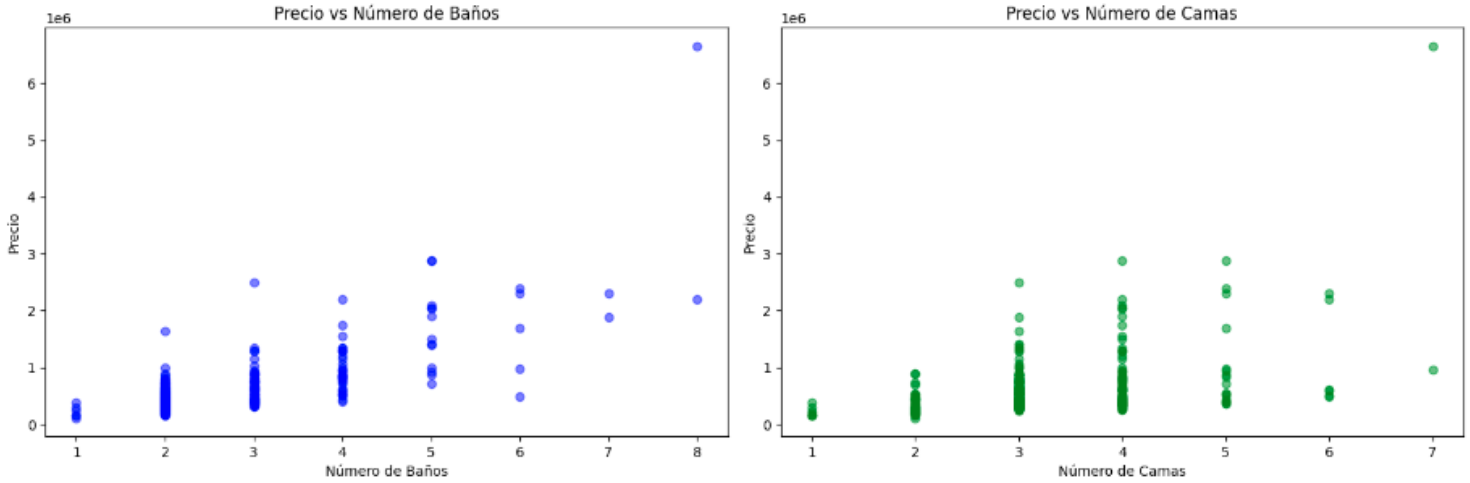
Los 5 códigos postales más caros con sus respectivas ciudades, estados y precios promedio:

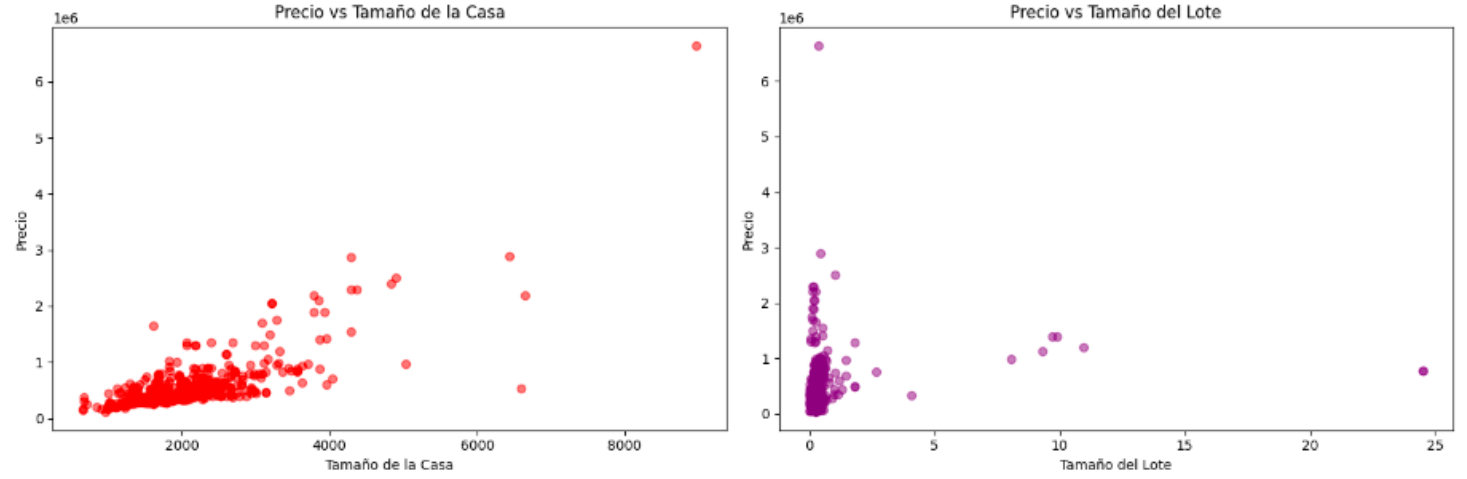
	zip_code	price	city	state
4105	10985.0	3.000000e+07	Thompson Ridge	New York
791	2672.0	1.810000e+07	Barnstable	Massachusetts
2453	6753.0	1.333000e+07	Cornwall	Connecticut
4821	11962.0	1.291905e+07	Sagaponack	New York
717	2543.0	1.230000e+07	Woods Hole	Massachusetts

Analysis of features respect to a single zip_code

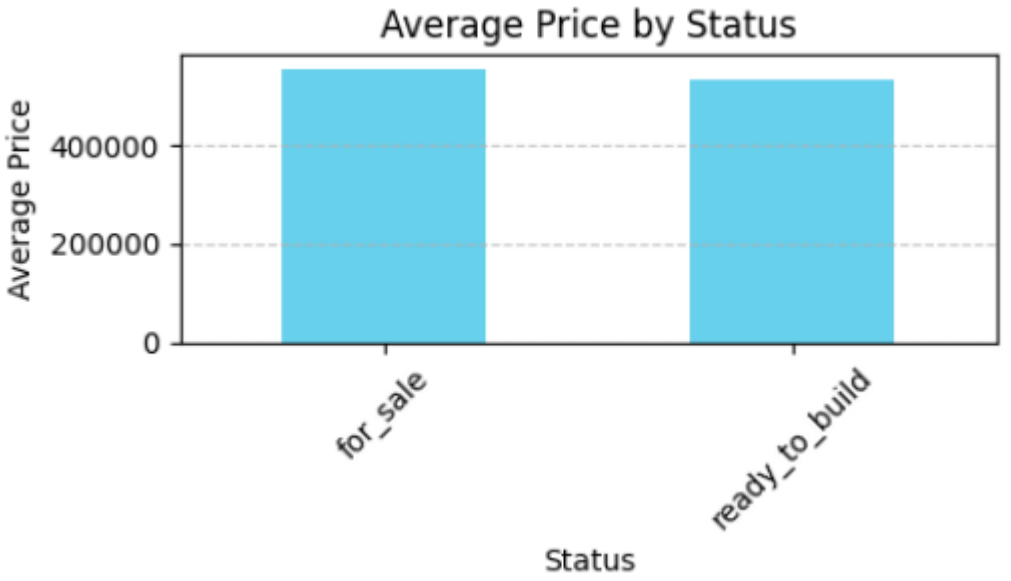
We can notice the linear growing of the price in proportion to the features.

The acre lot is not



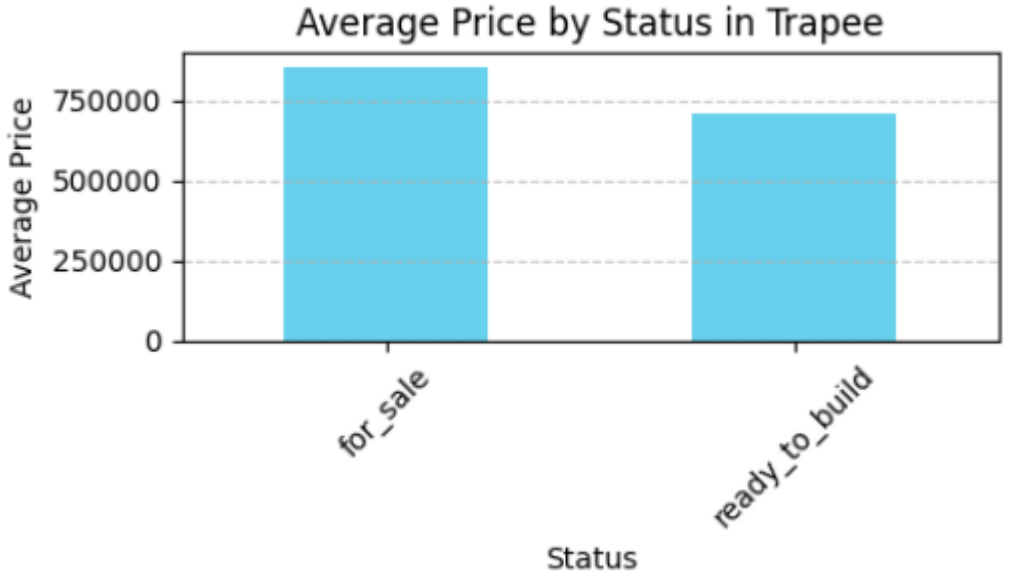


Analysis of features respect to a single zip_code



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?

Conclusion

Modelo Predictivo.

- INTELIGENCIA ARTIFICIAL A LA CIENCIA DE DATOS

Objetivos

- Price Prediction

- Predict the price of properties based on their characteristics.

Decisions.

Processing:
Drop duplicates
Drop Missing Values
Drop ‘state’ and ‘city’
Drop ‘prev_sold_date’
Change ‘status’ to categorical

Outliers:
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

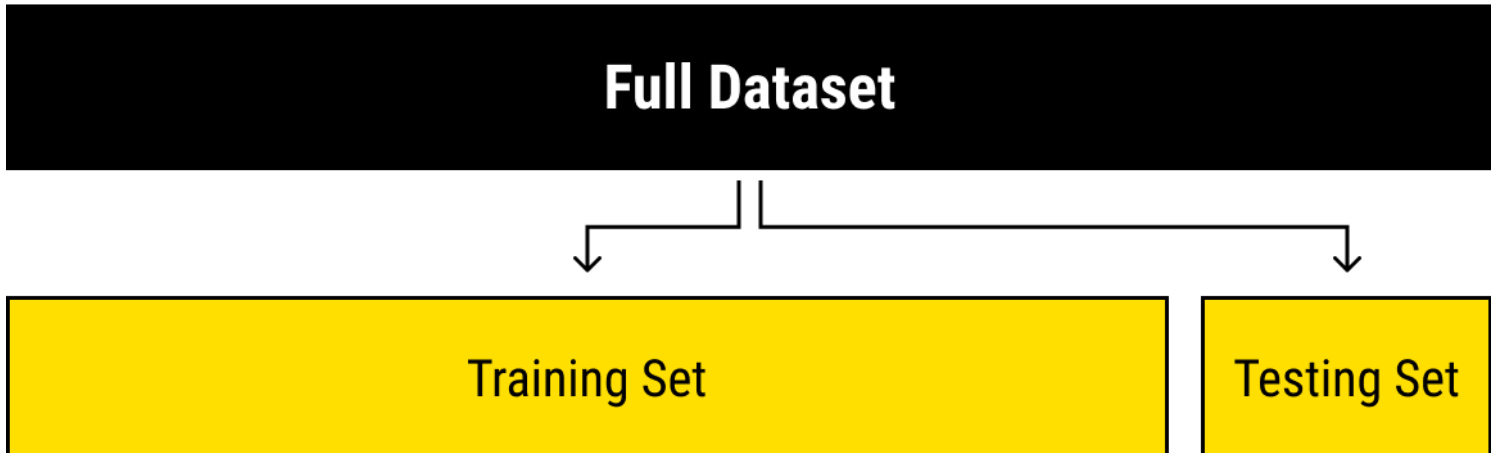
	status	bed	bath	acre_lot	city	state	zip_code
0	for_sale	3.0	2.0	0.12	Adjuntas	Puerto Rico	601.0
1	for_sale	4.0	2.0	0.08	Adjuntas	Puerto Rico	601.0
2	for_sale	2.0	1.0	0.15	Juana Diaz	Puerto Rico	795.0
3	for_sale	4.0	2.0	0.10	Ponce	Puerto Rico	731.0
4	for_sale	6.0	2.0	0.05	Mayaguez	Puerto Rico	680.0

	house_size	prev_sold_date	price
0	920.0	NaN	105000.0
1	1527.0	NaN	80000.0
2	748.0	NaN	67000.0
3	1800.0	NaN	145000.0
4	NaN	NaN	65000.0

3096565 rows

	status	bed	bath	acre_lot	zip_code	house_size	price
0	1	3	2	0.12	601	920	105000
1	1	4	2	0.08	601	1527	80000
2	1	2	1	0.15	795	748	67000
3	1	4	2	0.10	731	1800	145000
4	1	4	3	0.46	612	2520	179000
...	...	...	...	...	...	...	...

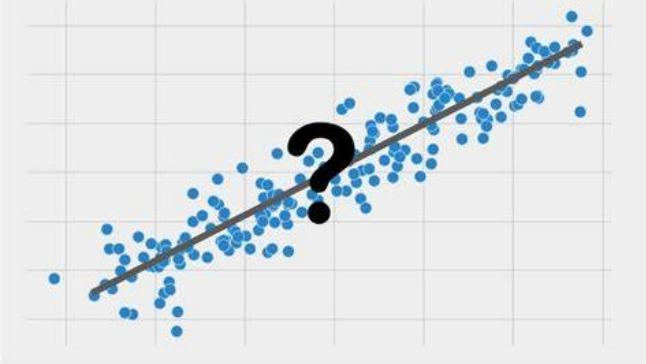
207480 rows



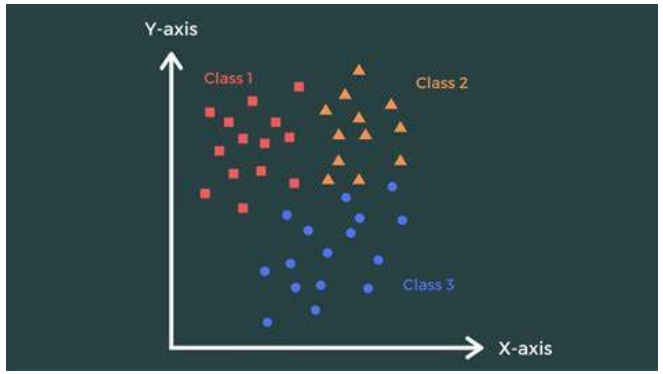
Split the data into training and testing sets (e.g., 80% training and 20% testing) to evaluate model performance.

Chosen Models

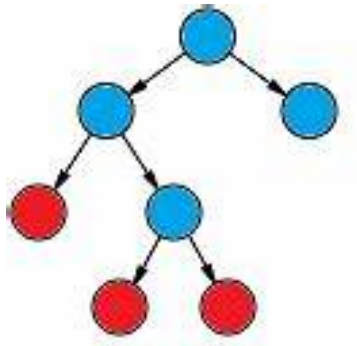
Linear Regression



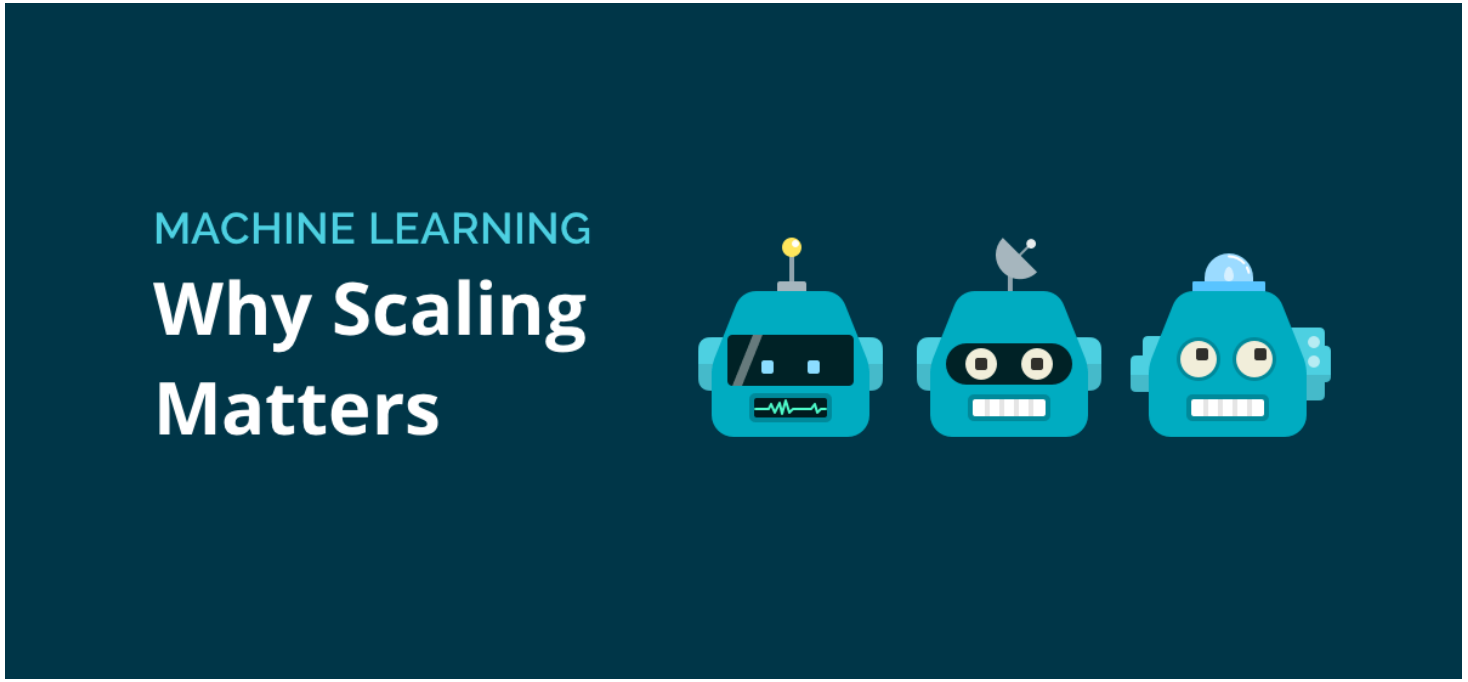
KNN



Random Forest

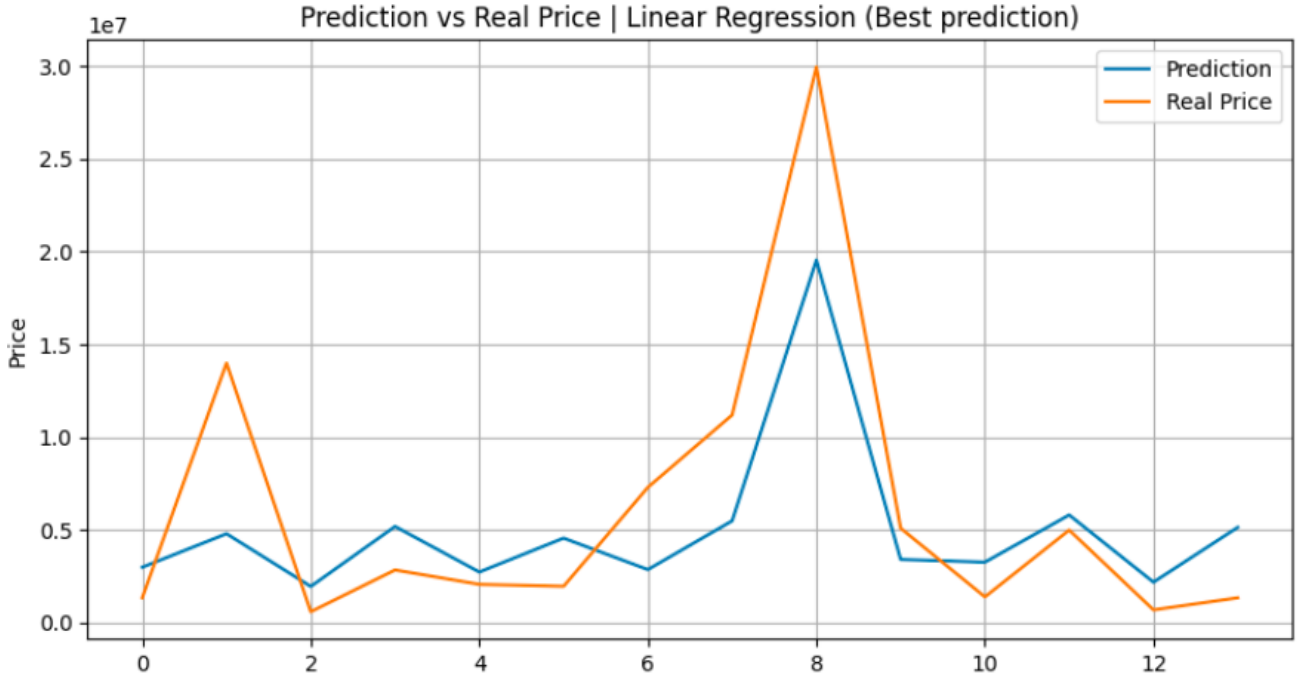


Variables were scaled to improve model performance by ensuring that variables contribute equitably.



Linear Regression

The data were separated according to the number of beds given the variety of property types.



Dataset 1
RMSE = 355821.12
R2 = 0.063

Dataset 2
RMSE = 702258.56
R2 = 0.271

Dataset 3
RMSE = 249168.24
R2 = 0.102

Dataset 4
RMSE = 515830.7
R2 = 0.086

Dataset 5
RMSE = 524257.23
R2 = 0.207

Dataset 6
RMSE = 1337970.42
R2 = 0.248

Dataset 7
RMSE = 2519518.14
R2 = 0.336

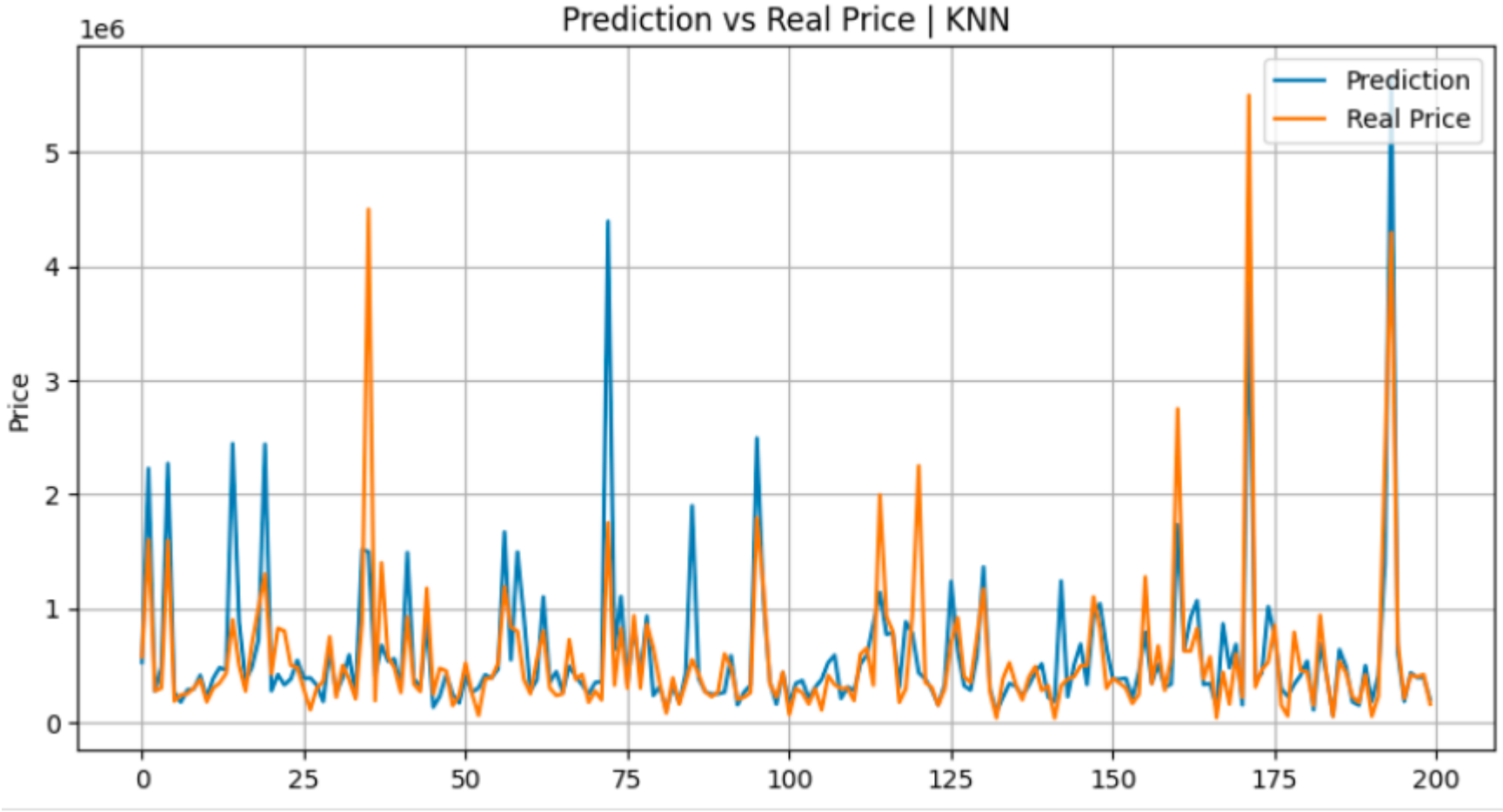
Dataset 8
RMSE = 2779417.06
R2 = 0.405

Dataset 9
RMSE = 5254063.17
R2 = 0.313

Dataset 10
RMSE = 4518202.15
R2 = 0.656

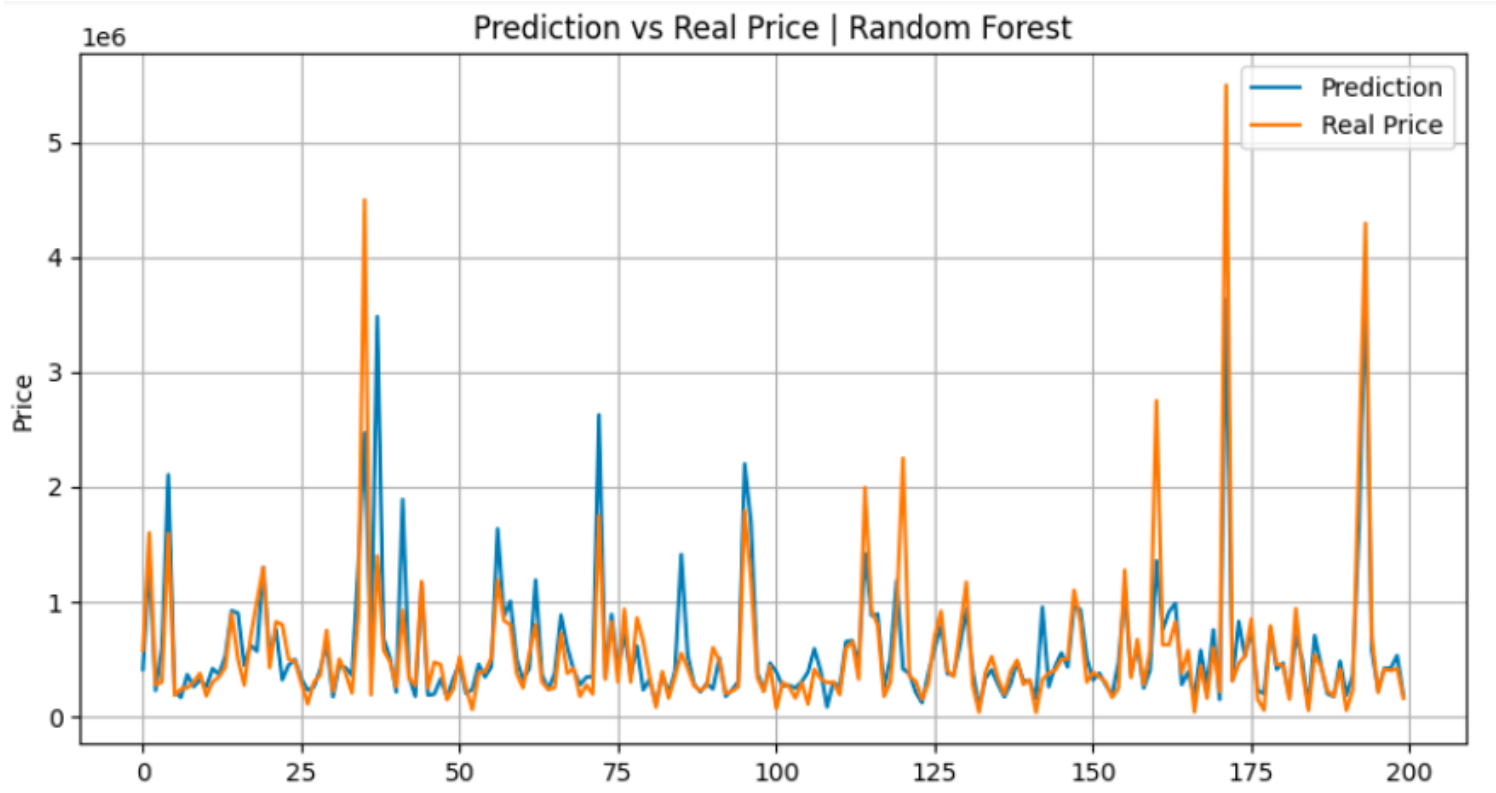
KNN

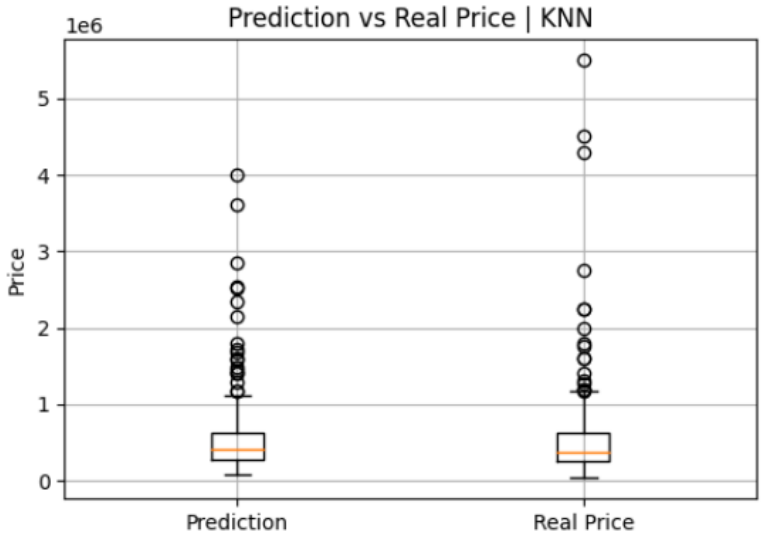
KNN (K-Nearest Neighbors) is a machine learning algorithm that predicts the classification or value of a new data point by considering the majority class or average value of its nearest neighbors in the feature space.



RMSE for KNN : 869620.71
R2 for KNN : 0.4518

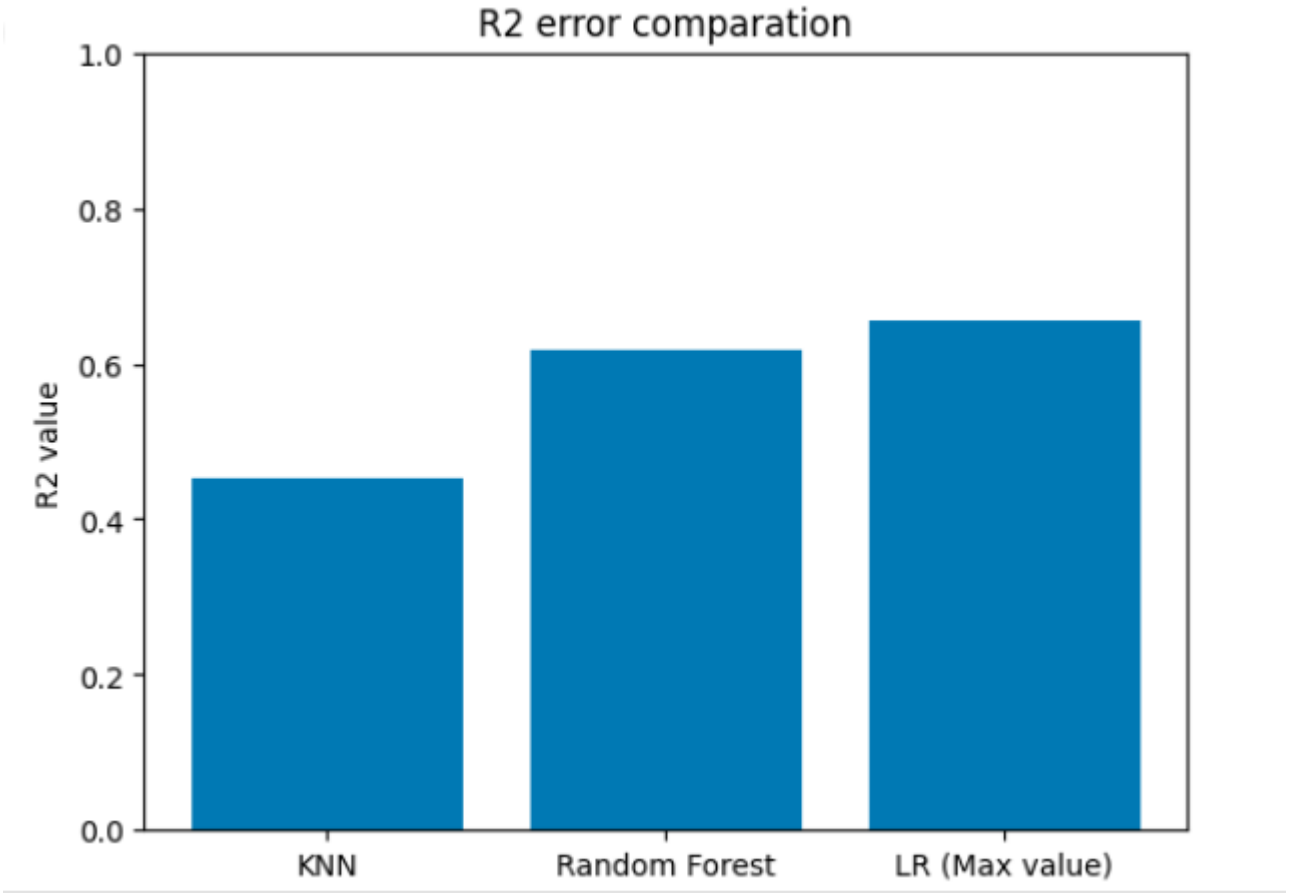
Random Forest





RMSE : 726526.68
R2 : 0.6174

Comparation



Interpretation

Random Forest

The best predictor model out of this research.

Would be good to re-adjust the importance of the variables.

Data

The types of property are very variable, it would be convenient to separate the data by type of property, hotels, personal homes, multi-family, etc.

Other models

It would be convenient to use a more advanced method such as Neural Networks

Due the great outliers in the data it is not easy to get a good prediction it would be better to use other IA techniques like Neural Networks.

Conclusion

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

-NEIL ARMSTRONG