Caso de estudio
Regresión

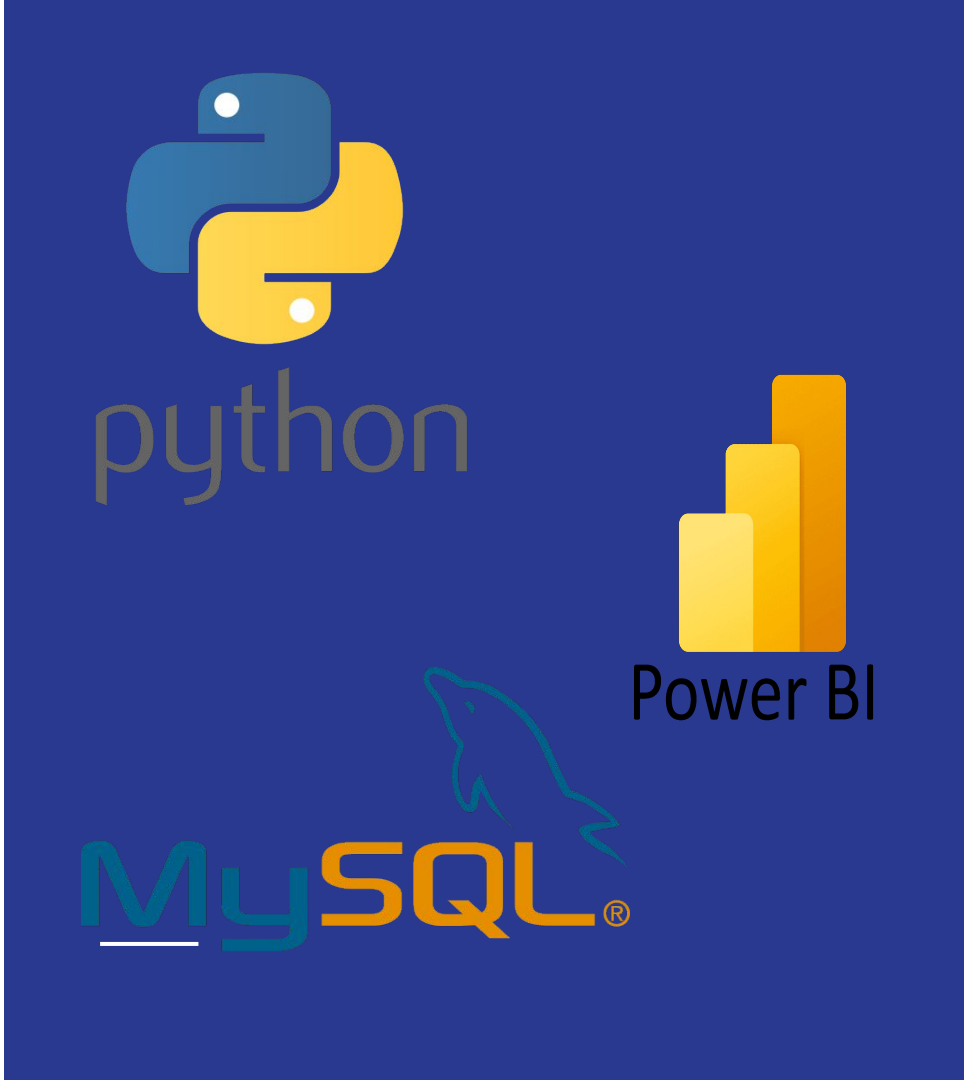
OBJETIVOS

- Construir un modelo que prediga el precio de una casa en función de las características proporcionadas en el conjunto de datos.

- Explorar las características de las viviendas.

- Uno de esos parámetros incluye comprender qué factores son responsables del mayor valor de la propiedad: $650 mil y más.

PROGRAMAS A UTILIZAR



METODOLOGÍA

1. Importar bibliotecas y cargar el conjunto de datos
2. Visión general del conjunto de datos
3. Limpieza de datos
4. Análisis exploratorio de datos (EDA)
5. Modelización de datos y examinar el cruzado de modelos
6. Conclusión

,k’t5rx

Bibliotecas

- Pandas
- Seaborn
- Matplotlib.pyplot
- Sklearn.model_selection
- Sklearn.linear_model
- Sklearn.metrics

2. Visión general del conjunto de datos

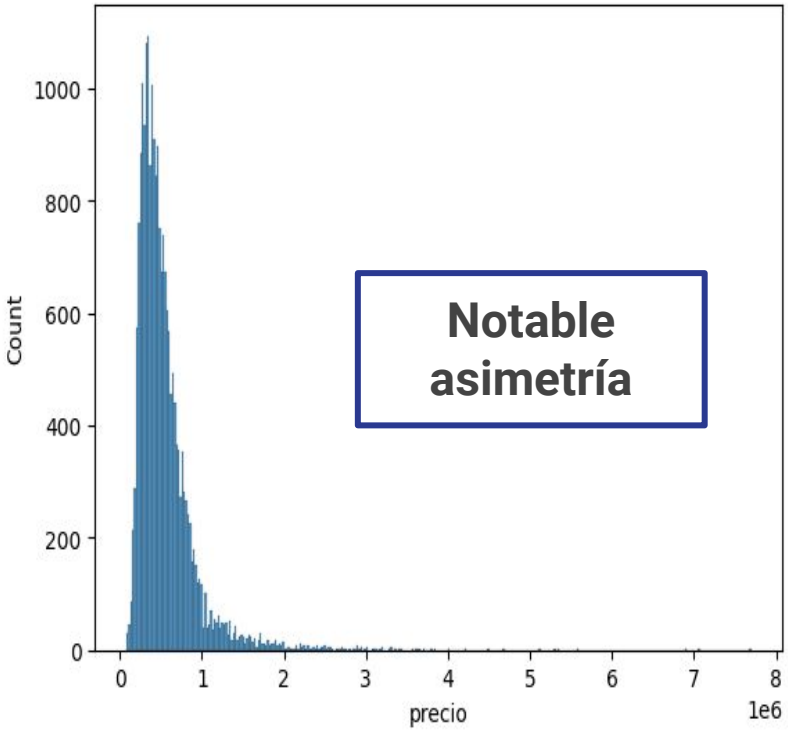
Datos de 21.597 propietarios

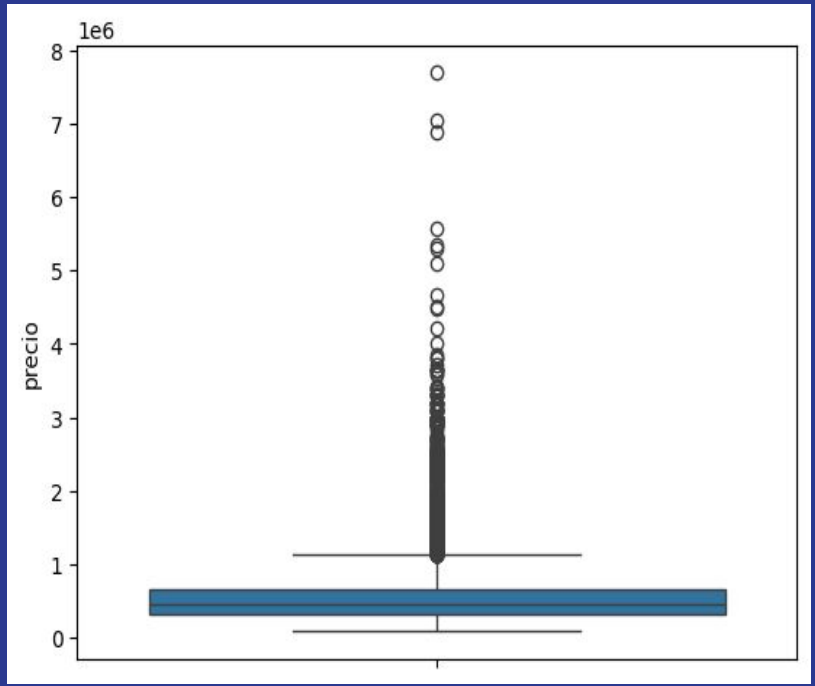
Ventas entre Mayo 2014 - Mayo 2015

3. Limpieza de datos

- Buscar NaNs.
- Cambiar nombres a Español.
- Dividir columnas categóricas y numéricas.
- Cambiar tipo de código postal string a un formato categórico.

4. Análisis exploratorio de datos (EDA)





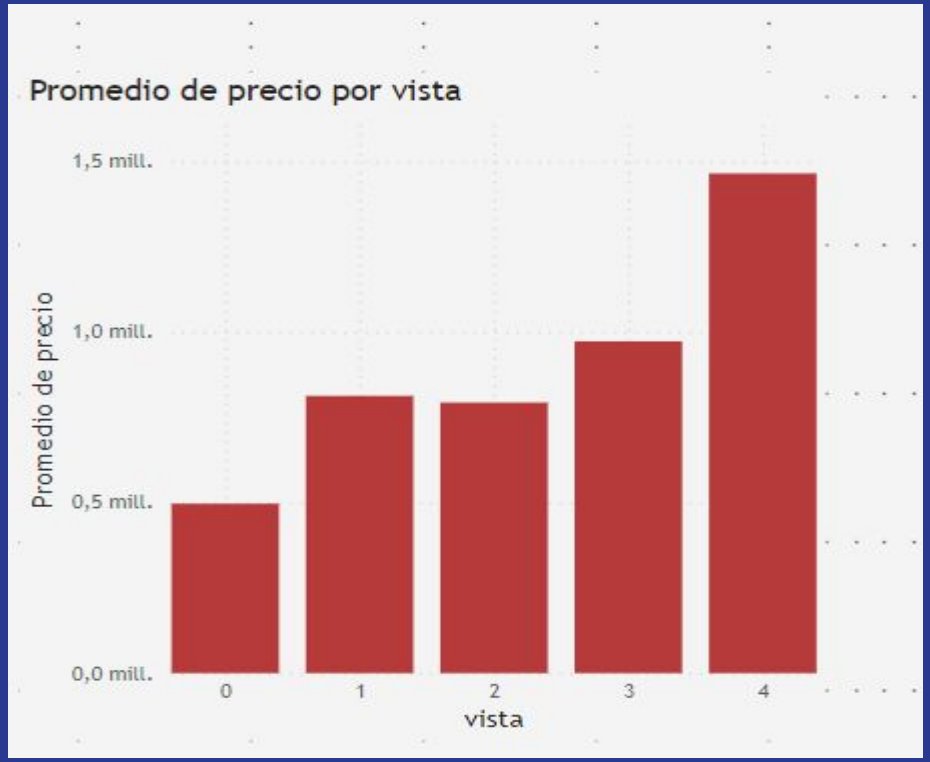
Columnas numéricas

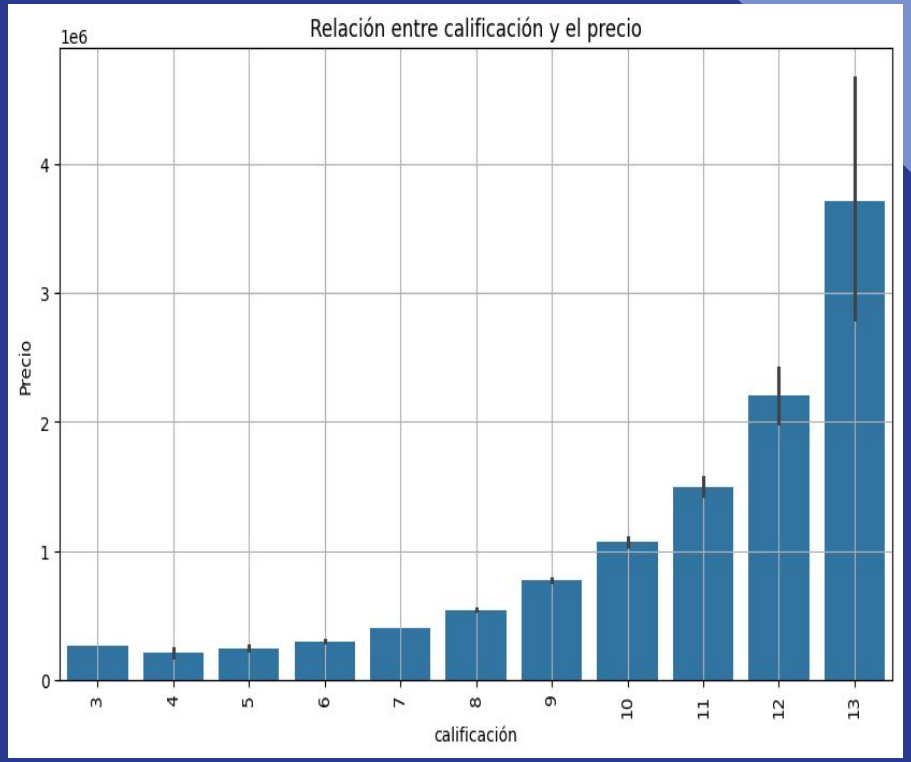
```
precio                        1.000000
pies_cuadrados_vivienda       0.701917
calificación                  0.667951
pies_cuadrados_sin_sotano     0.605368
pies_cuadrados_salon_15       0.585241
baños                         0.525906
vista                         0.397370
pies_cuadrados_sótano         0.323799
dormitorios                   0.308787
frente_al_mar                 0.266398
pisos                         0.256804
año_renovación                0.126424
pies_cuadrados_parcela        0.089876
pies_cuadrados_parcela_15     0.082845
año_construcción              0.053953
estado                        0.036056
código_postal                -0.053402
Name: precio, dtype: float64
```

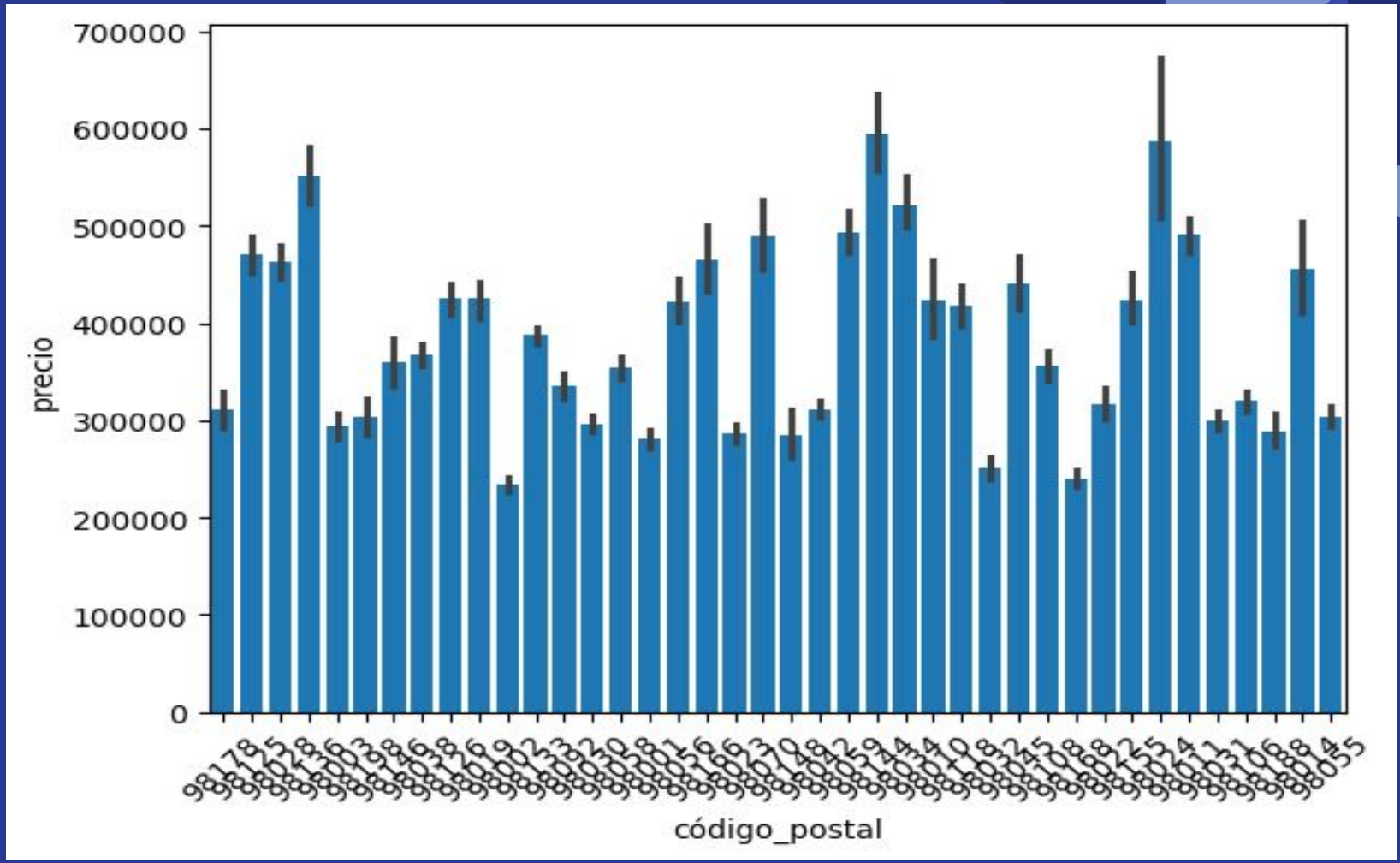
Datos más correlacionados

Las características en relación a las dimensiones y la clasificación es lo más relevante

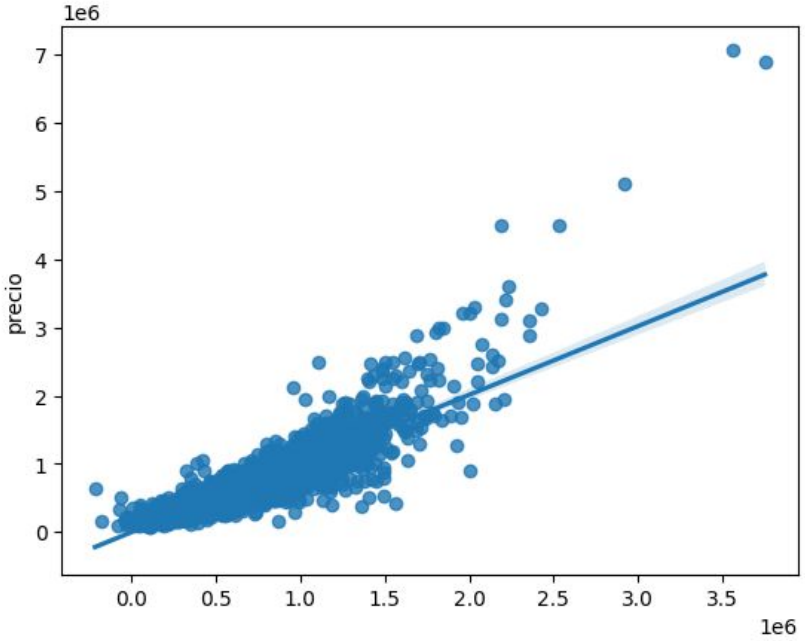
Columnas categóricas





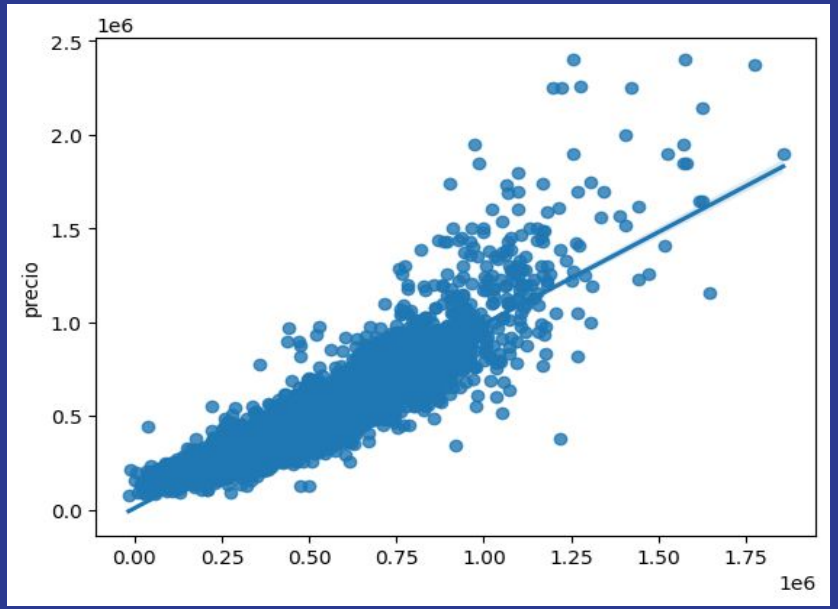

700000
600000
500000
400000
300000
200000
100000
0
precio
código_postal

5.MODELOS



Modelo de regresión lineal

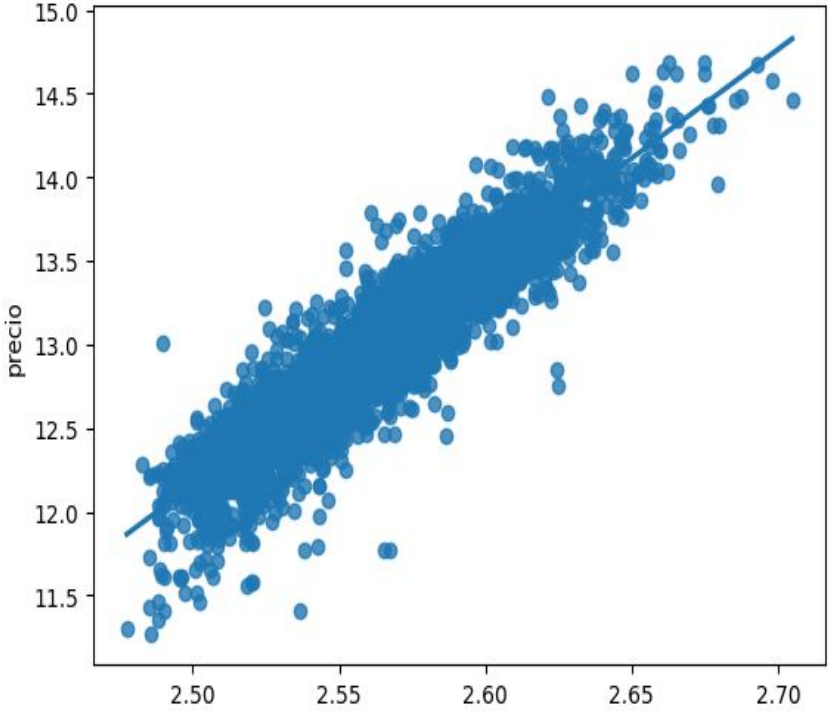
R2_score: 0.8
RMSE: 164236.79
MAE: 96394.52



Modelo de regresión lineal sin outliers

R2_score: 0.82
RMSE: 109831.76
MAE: 73258.02

5.MODELOS



MODELO LOGARÍTMICO SIN OUTLIERS

- R^2: 0.8661901132515291
- RMSE: 0.1746050194090754
- MAE: 0.127717827138633 78

Modelos >650K

Modelo de regresión lineal

R2_score: 0.8
RMSE: 164236.79
MAE: 96394.52

Modelo de regresión lineal sin outliers

R2_score: 0.84
RMSE: 83472.93
MAE: 61972.98

Modelo de logaritmico

R2_score: 0.81
RMSE: 87937.82
MAE: 60461.24

Modelo de KNN

R2_score: 0.39
RMSE: 159184.73
MAE: 123788.41

Modelo de Arbol de decisión:

R2_score: 0.59
RMSE: 130913.51
MAE: 90742.61

Modelo de Random Forest

R2_score: 0.79
RMSE: 92613.17
MAE: 63873.34

Modelo de Gradient Boosting

R2_score: 0.74
RMSE: 104146.96
MAE: 77002.30

6. Conclusiones:

- **Tamaño y precio**: El tamaño de la vivienda y características como el número de baños y clasificación influyen significativamente en el precio.
- **Mejor modelo**: El modelo logarítmico es el más preciso y consistente en general.
- **Mejor modelo para precios altos**: Para viviendas superiores a $650,000, el mejor modelo es sin outliers y sin transformación logarítmica.