PROYECTO 1

Análisis del precio de las viviendas de Barcelona

Descripción breve

Desarrollo de un análisis completo y un modelo predictivo para los precios de viviendas en Barcelona, utilizando datos extraídos del portal Fotocasa y aplicando técnicas de extracción, manipulación y análisis de datos, así como algoritmos de Machine Learning, para predecir los precios de las viviendas en función de diversas características

El código de Python utilizado para realizar este trabajo se encuentra en:
https://github.com/carlosarreguib/PROYECTO_1

Archivos relevantes:
- Barcelona_Fotocasa_HousingPrices.csv
- Pre-processing.ipynb
- MachineLearning.ipynb

Juan Pablo Delzo, Carlos Arregui, Alan Rocamora

Contenido

Lista de figuras 2
Lista de tablas 3
1. Análisis exploratorio de los datos 4
1.1. Análisis por atributo 5
1.1.1. Neighborhood 5
1.1.2. Real estate 5
1.1.3. Rooms 6
1.1.4. Bathroom 6
1.1.5. Lift 7
1.1.6. Terrace 8
1.1.7. Square meters 8
1.1.8. Square meters price 9
1.2. Eliminación de atípicos 9
2. Modelos de predicción 10
2.1. Regresión lineal simple 10
2.2. Lasso 11
2.3. Ridge 12
2.4. ElasticNet 12
2.5. K-nearest neighbor (KNN) 13
2.6. Random Forest 13
2.7. Gradient Boosting 14
3. Conclusiones 16

Lista de figuras

Figura 1. Distribución de tipos de viviendas por barrios 4
Figura 2. Matriz de correlación..... 4
Figura 3. Gráfico de barras con el número de viviendas en cada barrio y gráfico de caja para los precios de los alquileres..... 5
Figura 4. Gráfico de barras con el número de cada tipo de viviendas y gráfico de caja para los precios de los alquileres..... 6
Figura 5. Gráfico de barras con el número de habitaciones y gráfico de caja para los precios de los alquileres..... 6
Figura 6. Gráfico de barras con el número de baños y gráfico de caja para los precios de los alquileres 7
Figura 7. Gráfico de barras con los pisos que disponen o no de ascensor y gráfico de caja para los precios de los alquileres 7
Figura 8. Gráfico de barras con los pisos que disponen o no de terraza y gráfico de caja para los precios de los alquileres..... 8
Figura 9. Distribución del número de viviendas en función de los metros cuadrados y gráfico de dispersión con todos los precios de las viviendas según sus metros cuadrados 8
Figura 10. Distribución del número de viviendas en función del precio por metro cuadrado y gráfico de dispersión con todos los precios de las viviendas según su precio por metro cuadrado..... 9
Figura 11. Distribución de los datos originales del número de viviendas según el precio del alquiler y distribución con transformación logarítmica 10
Figura 12. Gráfico del ajuste de Regresión Lineal Simple a los datos originales y a los datos con transformación logarítmica 11
Figura 13. Gráfico del ajuste de Lasso a los datos originales y a los datos con transformación logarítmica..... 11
Figura 14. Gráfico del ajuste de Ridge a los datos originales y a los datos con transformación logarítmica..... 12
Figura 15. Gráfico del ajuste de ElasticNet a los datos originales y a los datos con transformación logarítmica..... 12
Figura 16. Gráfico del ajuste de KNN a los datos originales y a los datos con transformación logarítmica..... 13
Figura 17. Gráfico del ajuste de Random Forest a los datos originales y a los datos con transformación logarítmica 14
Figura 18. Gráfico del ajuste de Gradient Boosting a los datos originales y a los datos con transformación logarítmica 14

Lista de tablas

Tabla 1. Resumen de resultados de todos los modelos 15

1. Análisis exploratorio de los datos

El análisis exploratorio realizado aporta una visión global de la situación del mercado inmobiliario de Barcelona. Los datos proporcionados contienen la siguiente información de un total de 8188 viviendas:

- Precio
- Número de habitaciones
- Número de baños
- Ascensor en finca
- Tipo de vivienda
- Barrio
- Precio por metro cuadrado

Para empezar, se grafica el total de cada tipo de viviendas hay por cada barrio. Queda evidenciado que la vivienda modal en Barcelona es la del tipo “flat” en todos los barrios. Como el conjunto de datos no es demasiado extenso y el resto de los valores de la fila son de utilidad, se decide que los valores nulos de la columna se rellenan con este valor categórico.

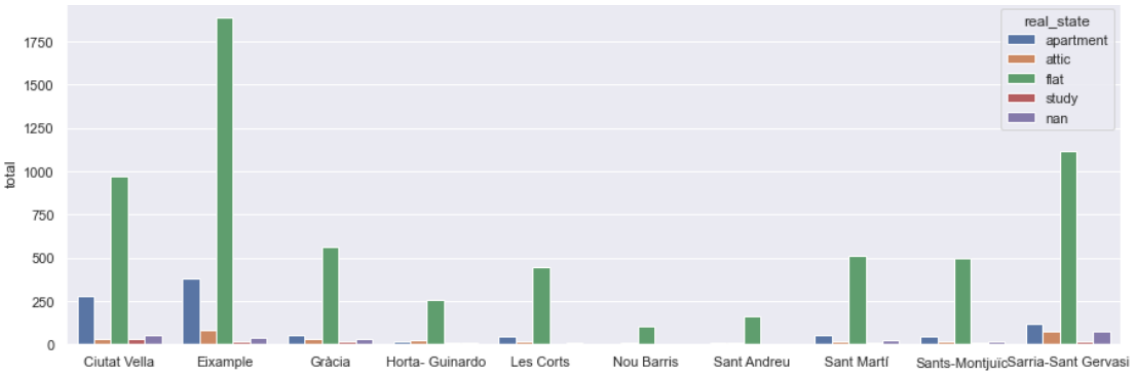


Figura 1. Distribución de tipos de viviendas por barrios

Para medir y visualizar la relación entre varias variables del conjunto de datos, se grafica la matriz de correlación eliminando el precio de la vivienda.

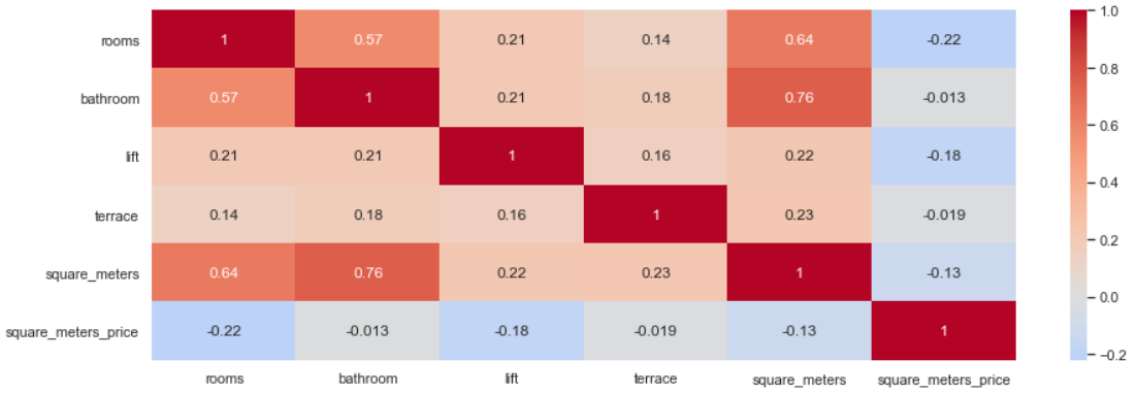


Figura 2. Matriz de correlación

Se observa que el coeficiente de correlación entre square_meters y bathroom es alto. Por tanto, se puede excluir bathroom en el diseño de modelos de predicción. Se debe de omitir el análisis a los atributos que provienen de boleanas.

En el análisis exploratorio se ha realizado un análisis por atributo. El objetivo es graficar a cada atributo respecto al precio.

1.1. Análisis por atributo

1.1.1. Neighborhood

El coeficiente de correlación entre barrio y precio es 0.09. De acuerdo con la correlación, se puede afirmar que no existe diferencia del precio de la vivienda por distrito.

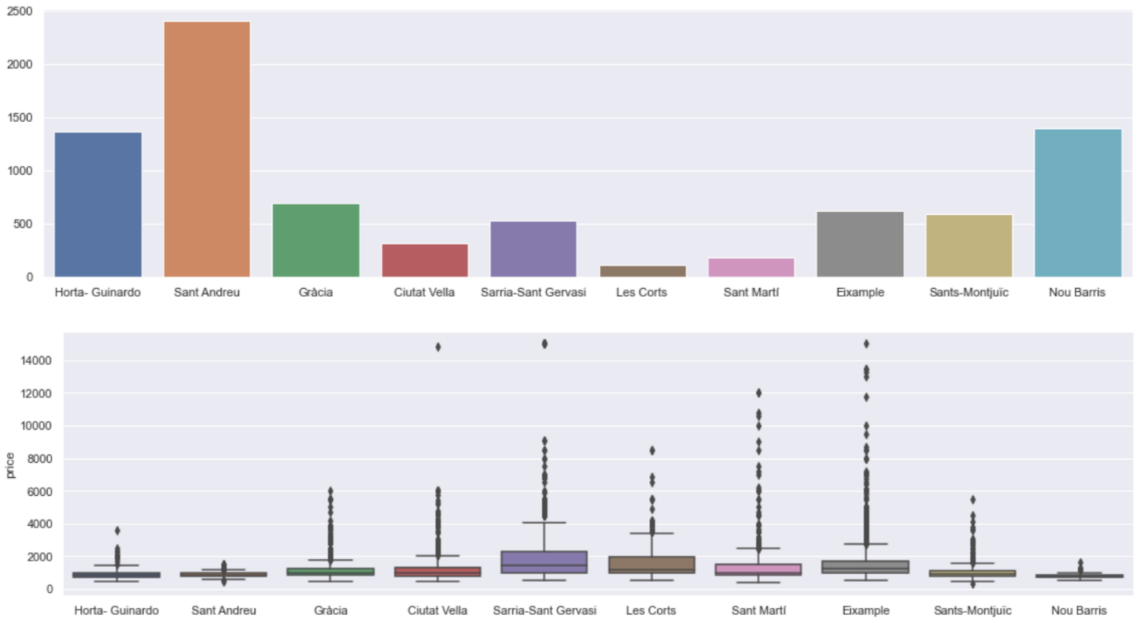
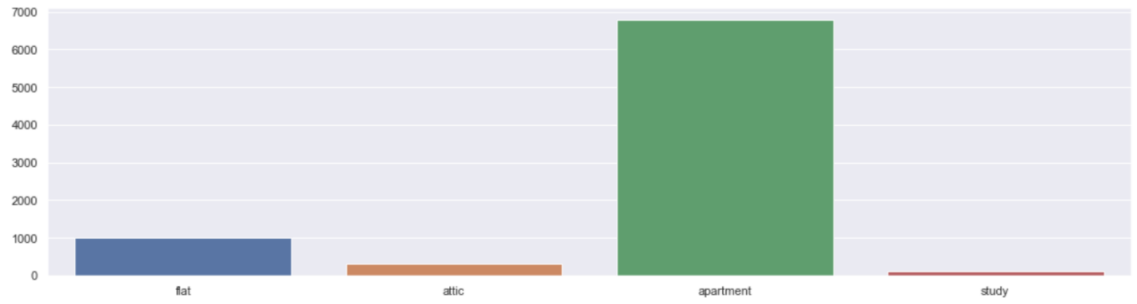


Figura 3. Gráfico de barras con el número de viviendas en cada barrio y gráfico de caja para los precios de los alquileres

1.1.2. Real estate

El coeficiente de correlación entre el tipo de vivienda y el precio es -0.25. Se puede descartar que no existe una relación entre el precio y tipo de vivienda.



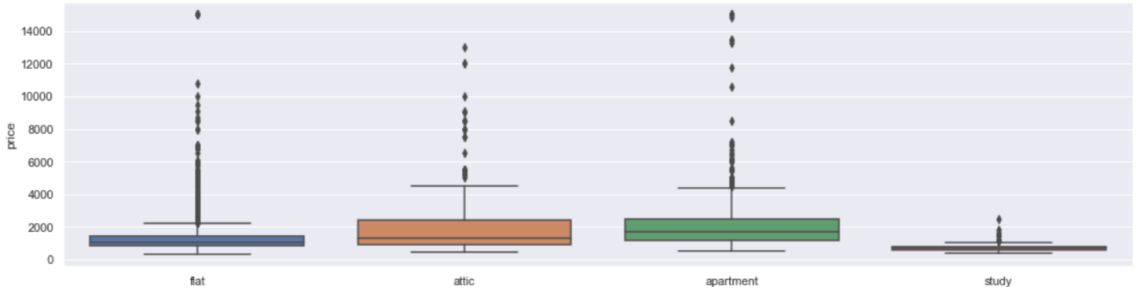


Figura 4. Gráfico de barras con el número de cada tipo de viviendas y gráfico de caja para los precios de los alquileres

1.1.3. Rooms

El coeficiente de correlación entre el número de habitaciones y el precio es de 0.35. No se puede descartar que no haya una relación entre el número de habitaciones y el precio, y los valores que prevalecen son 2 y 3 habitaciones.

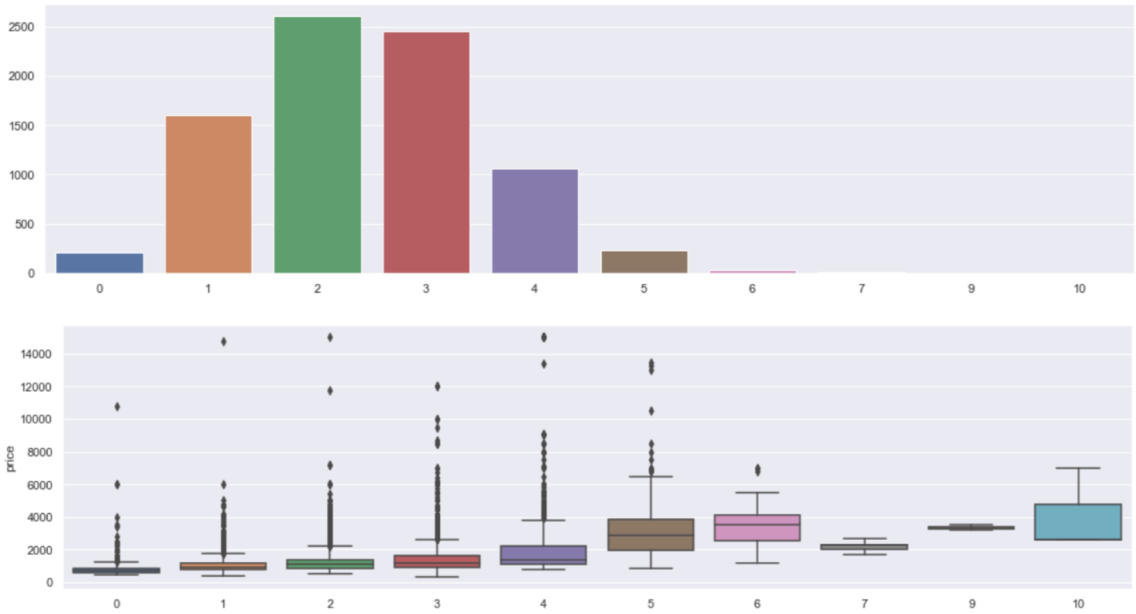


Figura 5. Gráfico de barras con el número de habitaciones y gráfico de caja para los precios de los alquileres

1.1.4. Bathroom

El coeficiente de correlación entre el número de baños y el precio de la vivienda es de 0.58. Es un coeficiente mayor que el de las habitaciones, con la moda del atributo en 1 baño

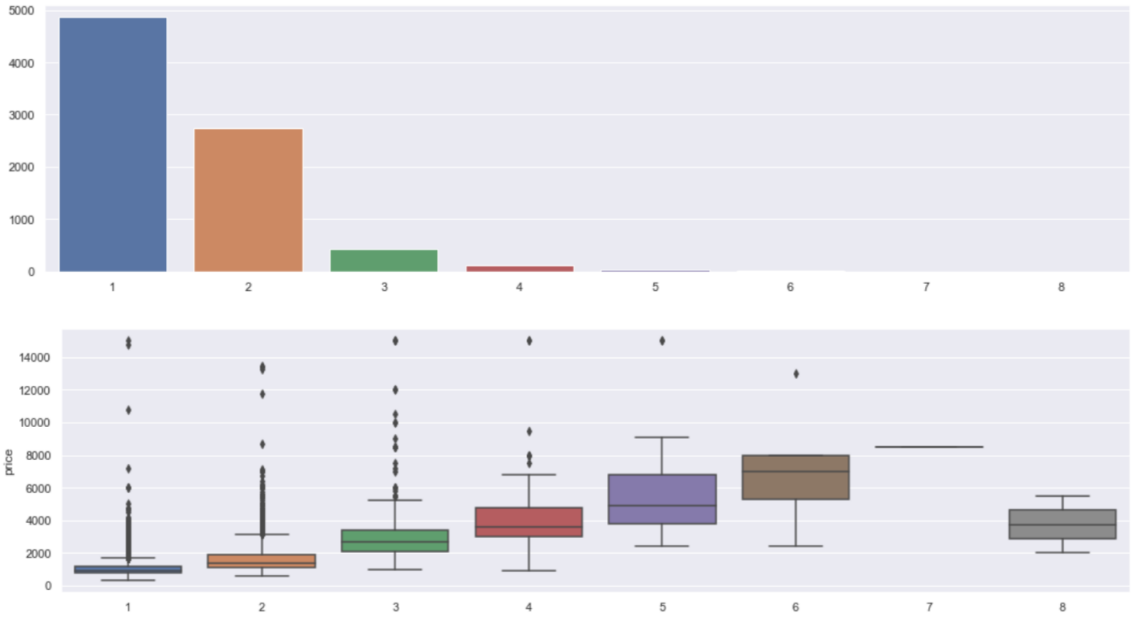


Figura 6. Gráfico de barras con el número de baños y gráfico de caja para los precios de los alquileres

1.1.5. Lift

El coeficiente de correlación entre el ascensor y el precio de la vivienda es 0.06. Se comprueba que la mayoría de las fincas en Barcelona disponen de ascensor, aunque no afecte en casi ninguna medida al precio.

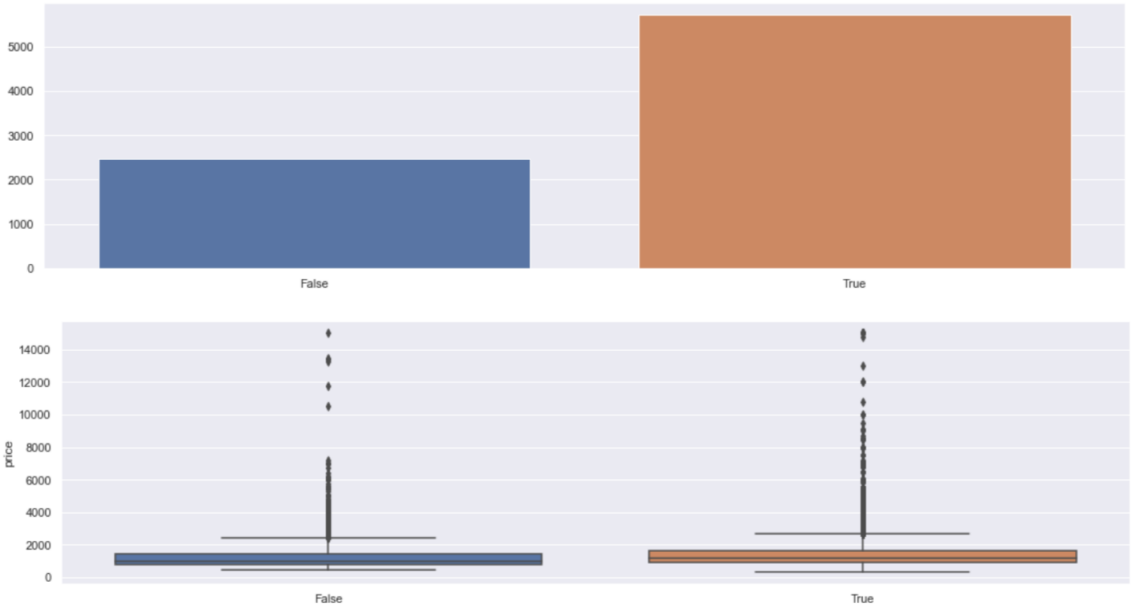


Figura 7. Gráfico de barras con los pisos que disponen o no de ascensor y gráfico de caja para los precios de los alquileres

1.1.6. Terrace

El coeficiente de correlación entre la terraza y el precio de la vivienda es 0.17. La mayoría de pisos no disponen de terraza.

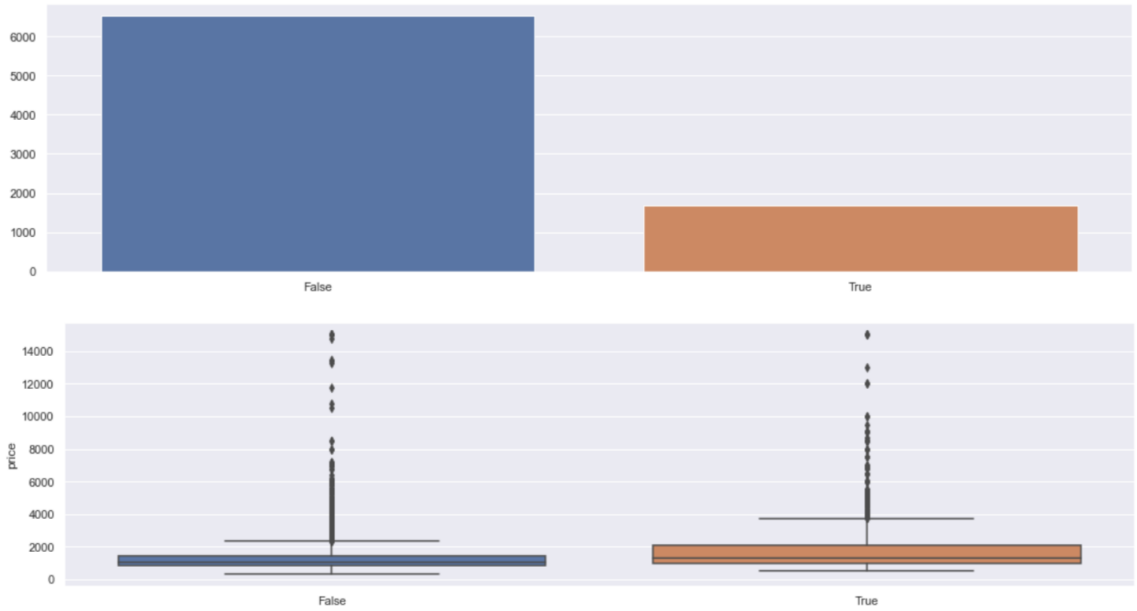


Figura 8. Gráfico de barras con los pisos que disponen o no de terraza y gráfico de caja para los precios de los alquileres

1.1.7. Square meters

El coeficiente de correlación entre los metros cuadrados del piso y el precio es 0.69. Hay una alta correlación y la distribución se ajusta a una campana de Gauss.

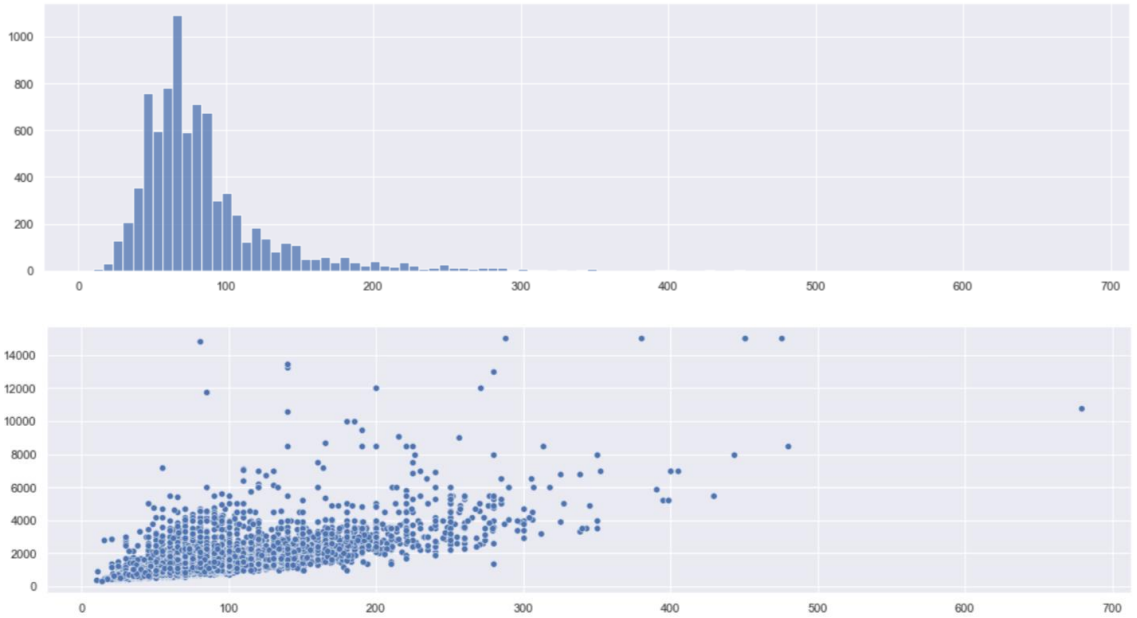


Figura 9. Distribución del número de viviendas en función de los metros cuadrados y gráfico de dispersión con todos los precios de las viviendas según sus metros cuadrados

1.1.8. Square meters price

El coeficiente de correlación entre el precio por metro cuadrado y el precio de la vivienda es 0.58, menor que respecto a los metros cuadrados. El histograma es claramente asimétrico desplazado hacia los precios bajos.

Figura 10. Distribución del número de viviendas en función del precio por metro cuadrado y gráfico de dispersión con todos los precios de las viviendas según su precio por metro cuadrado

1.2. Eliminación de atípicos

Se detectan varios atípicos, que se eliminan del dataframe original. Las filas seleccionadas son: 6951, 1772, 2427, 2754, 4220, 4750, 7646, 7928.

2. Modelos de predicción

La distribución de los precios de la vivienda tiene un sesgo positivo muy evidente. Se aplica una transformación logarítmica para eliminar el sesgo y aproximar más la distribución a una distribución normal.

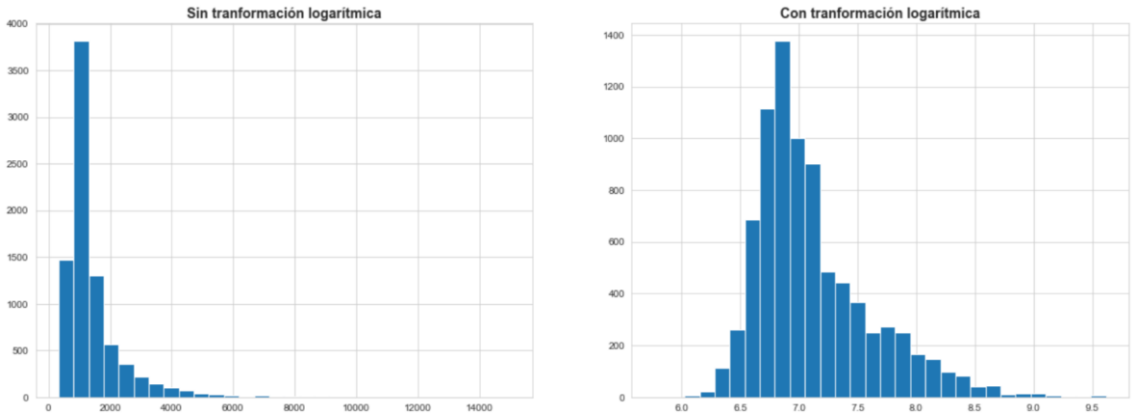


Figura 11. Distribución de los datos originales del número de viviendas según el precio del alquiler y distribución con transformación logarítmica

Se han aplicado los modelos de predicción a ambas distribuciones para comparar si algún modelo se ve mejorado. Los modelos aplicados para comparar los errores y los tiempos de procesamiento para ver cuál es el que predice mejores resultados han sido los siguientes:

- Regresión lineal simple
- Lasso
- Ridge
- ElasticNet
- K-nearest neighbor
- Random Forest
- Gradient Boosting

En todos los ajustes se ha utilizado una fracción del 80% de los datos para entrenar el modelo y un 20% para testear.

Para enseñar el ajuste de todos los modelos, se presentan todos los gráficos de todos los modelos para la distribución original y para la distribución con transformación logarítmica.

2.1. Regresión lineal simple

La regresión lineal simple es un modelo estadístico que describe la relación entre dos variables, una independiente y una dependiente, mediante una línea recta. Su objetivo es predecir el valor de la variable dependiente en función de la independiente, encontrando la mejor línea que minimice las diferencias entre los valores observados y los valores predichos por el modelo.

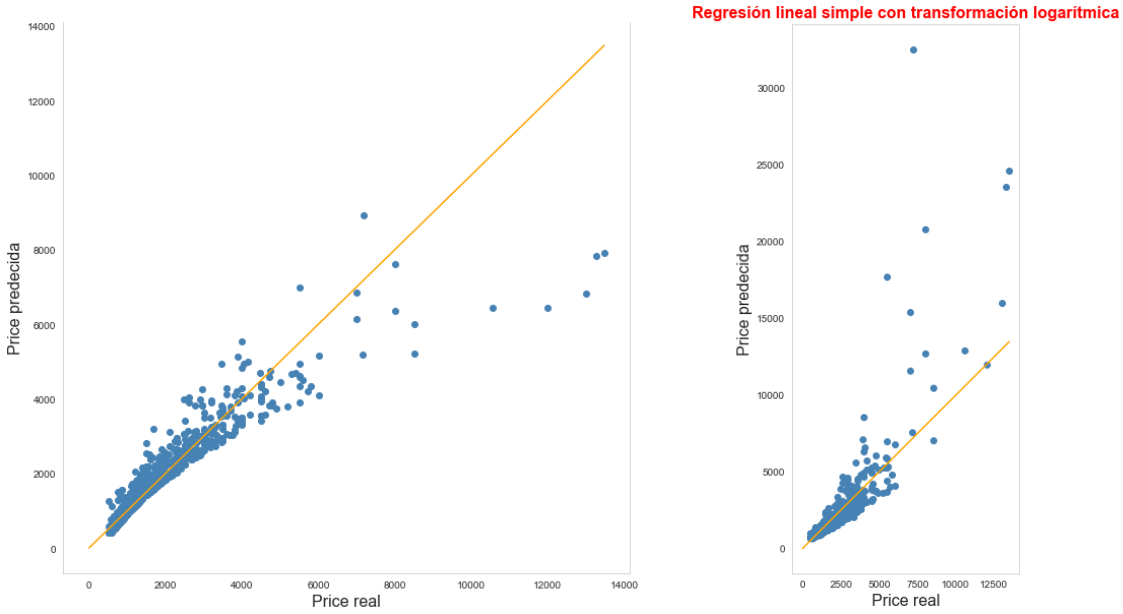


Figura 12. Gráfico del ajuste de Regresión Lineal Simple a los datos originales y a los datos con transformación logarítmica

2.2. Lasso

Lasso (*Least Absolute Shrinkage and Selection Operator*) es un modelo de regresión lineal que realiza selección de variables y regularización para mejorar la precisión de predicciones. Al agregar una penalización basada en el valor absoluto de los coeficientes, Lasso reduce algunos coeficientes a cero, eliminando variables menos relevantes y generando un modelo más simple y eficiente.

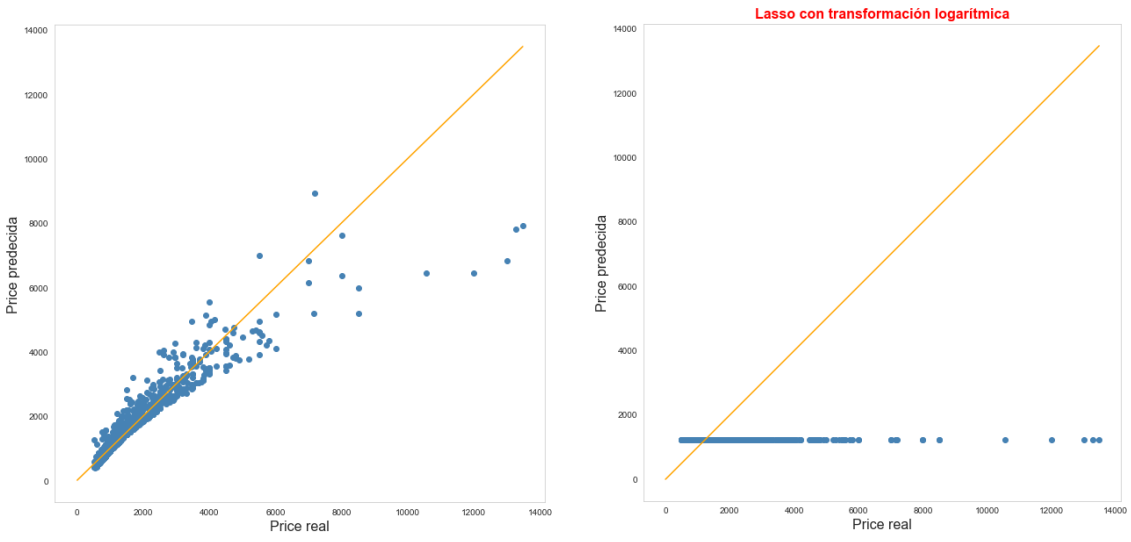


Figura 13. Gráfico del ajuste de Lasso a los datos originales y a los datos con transformación logarítmica

2.3. Ridge

Ridge es un modelo de regresión lineal que utiliza una técnica de regularización para reducir el sobreajuste. Añade una penalización basada en el cuadrado de los coeficientes de las variables, lo que evita que algunos de ellos crezcan demasiado y, en cambio, los mantiene pequeños. Esto ayuda a mejorar la precisión del modelo, especialmente cuando hay alta colinealidad entre variables independientes.

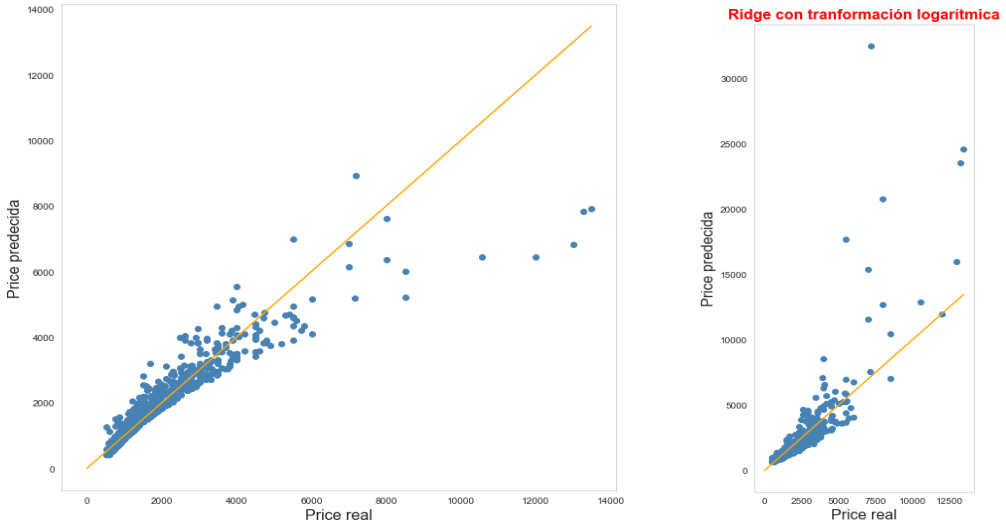


Figura 14. Gráfico del ajuste de Ridge a los datos originales y a los datos con transformación logarítmica

2.4. ElasticNet

ElasticNet es un modelo de regresión lineal que combina las penalizaciones de Lasso y Ridge para lograr tanto selección de variables como regularización. Utiliza una combinación de los valores absolutos y cuadrados de los coeficientes, lo que permite manejar mejor conjuntos de datos con muchas variables correlacionadas, mejorando la precisión y la interpretabilidad del modelo.

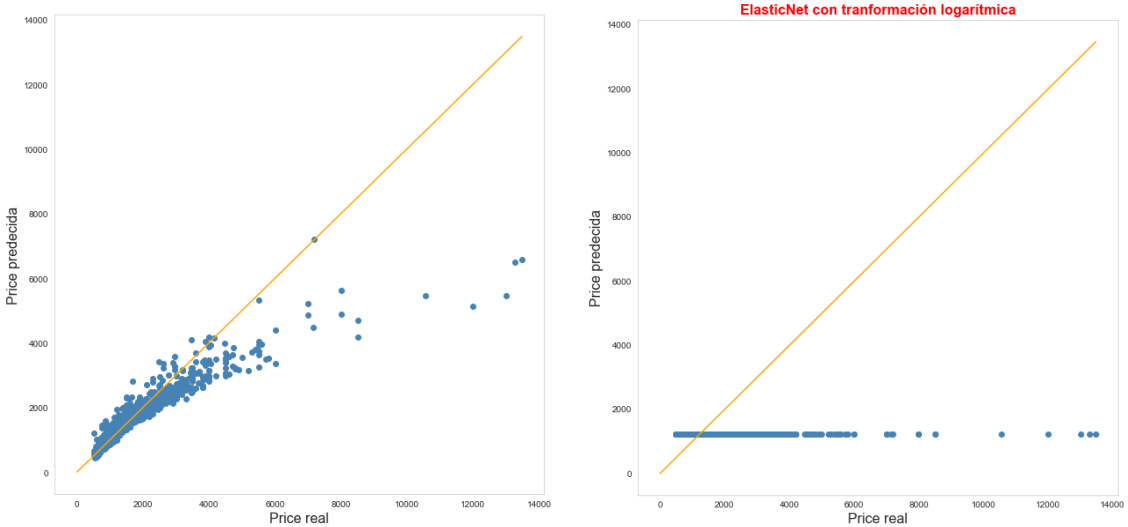


Figura 15. Gráfico del ajuste de ElasticNet a los datos originales y a los datos con transformación logarítmica

2.5. K-nearest neighbor (KNN)

K-Nearest Neighbors (KNN) es un modelo de aprendizaje supervisado que clasifica un dato nuevo en función de su proximidad a los datos existentes. La clasificación se determina según las "K" observaciones más cercanas (vecinos) en el conjunto de datos, asignando la categoría mayoritaria entre estos vecinos al dato nuevo. KNN es simple y eficaz, especialmente en problemas de clasificación y regresión con conjuntos de datos pequeños y bien distribuidos.

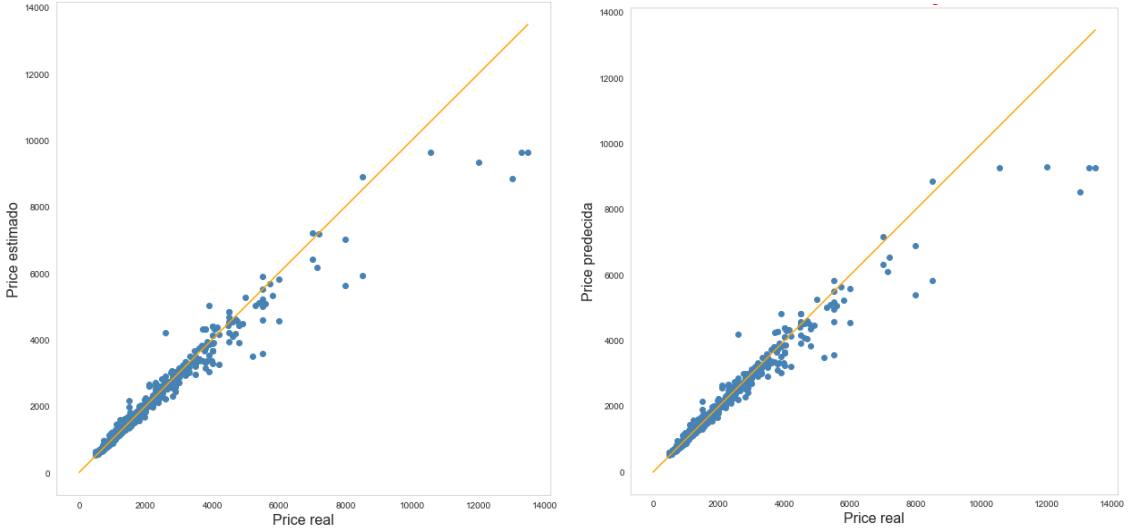


Figura 16. Gráfico del ajuste de KNN a los datos originales y a los datos con transformación logarítmica

2.6. Random Forest

Random Forest es un modelo de aprendizaje supervisado que combina múltiples árboles de decisión para mejorar la precisión y reducir el sobreajuste. Cada árbol en el bosque se entrena con una muestra aleatoria del conjunto de datos y realiza una predicción; luego, el modelo final elige la predicción promedio (para regresión) o la mayoría de los votos (para clasificación) entre todos los árboles. Esto permite a Random Forest manejar datos complejos y reducir la varianza en las predicciones.

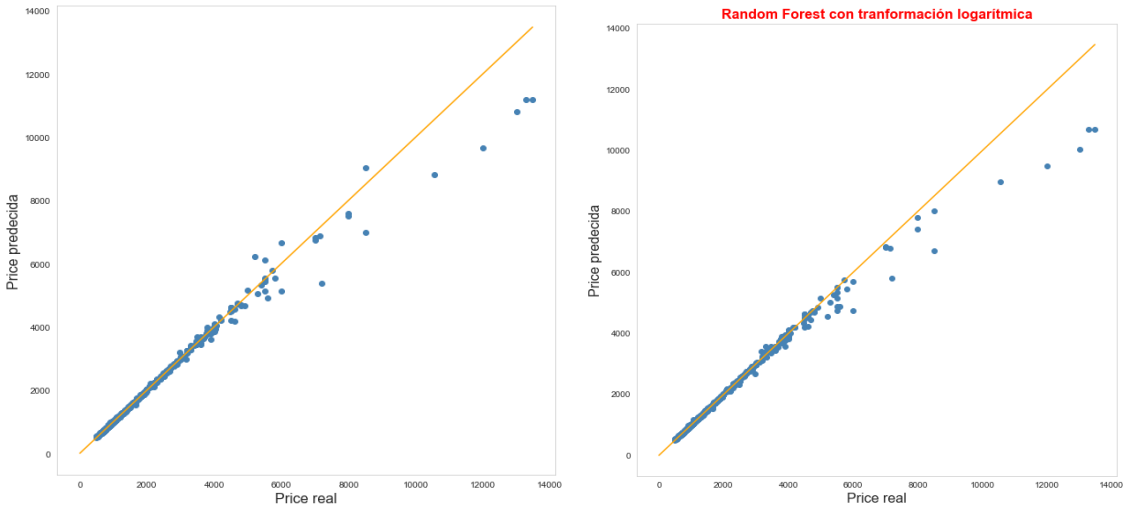


Figura 17. Gráfico del ajuste de Random Forest a los datos originales y a los datos con transformación logarítmica

2.7. Gradient Boosting

Gradient Boosting es un modelo de aprendizaje supervisado que crea una serie de árboles de decisión de forma secuencial, donde cada árbol nuevo corrige los errores del árbol anterior. Utiliza el gradiente del error para ajustar cada árbol sucesivo, de modo que el modelo mejore progresivamente. Este enfoque permite que Gradient Boosting sea preciso y eficaz en tareas de clasificación y regresión, aunque requiere una buena configuración para evitar el sobreajuste.

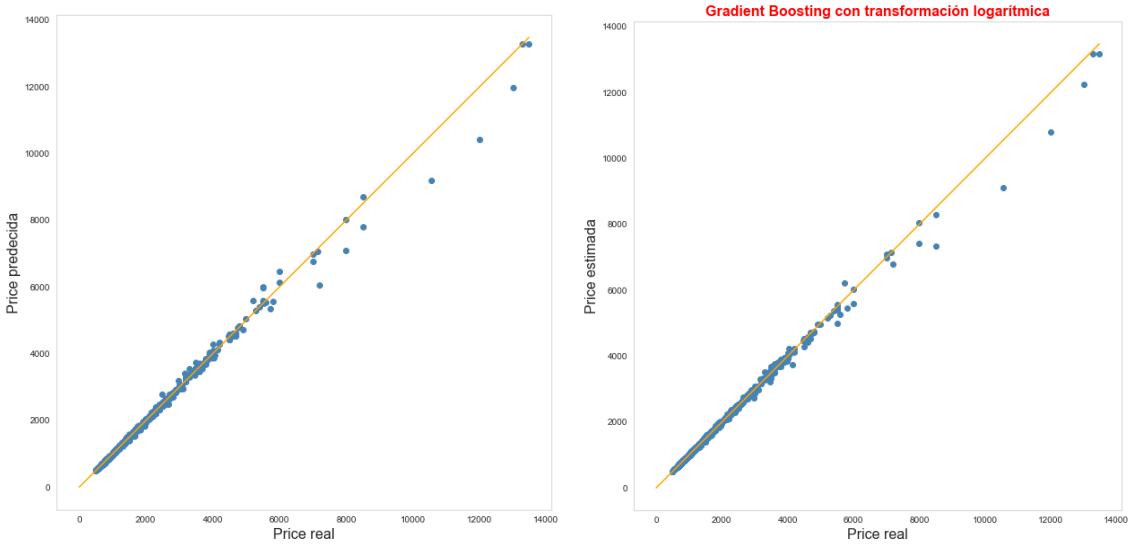


Figura 18. Gráfico del ajuste de Gradient Boosting a los datos originales y a los datos con transformación logarítmica

A continuación, se presenta una tabla resumen de los resultados de los ajustes de todos los modelos para las dos distribuciones mencionadas más arriba.

Tabla 1. Resumen de resultados de todos los modelos

Modelo	Error distribución original	Error distribución transformación logarítmica	Tiempo de procesamiento	
			Original	Transf. log.
Regresión lineal simple	165.17 €	256.03 €	2.12 s	927 ms
Lasso	164.90 €	612.21 €	87.6 ms	47.1 ms
Ridge	223.51 €	256.00 €	42.3 ms	51.9 ms
ElasticNet	165.15 €	612.21 €	37.4 ms	49.6 ms
K-nearest neighbor	68.63 €	70.52 €	426 ms	411 ms
Random Forest	22.27 €	24.11 €	12.6 s	13.1 s
Gradient Boosting	22.62 €	20.53 €	10.3 s	9.92 s

Queda evidente que la transformación logarítmica únicamente ayuda en el modelo Gradient Boosting, donde el error se disminuye en 2.09 € respecto a la distribución original.

Por otro lado, es importante fijarse en el tiempo de procesamiento de cada modelo. Los modelos más simples son los de Regresión Lineal Simple, Lasso, Ridge y ElasticNet. Todos excepto la Regresión Lineal Simple, presentan tiempos de procesamiento inferiores a un segundo. Sin embargo, los modelos de Random Forest y Gradient Boosting tienen un mayor tiempo de procesamiento porque construyen múltiples árboles de decisión, lo que requiere dividir los datos repetidamente. En Gradient Boosting, los árboles se construyen secuencialmente, aumentando el tiempo debido a la dependencia entre ellos. Además, la optimización de hiperparámetros en ambos modelos agrega más carga computacional, especialmente con grandes conjuntos de datos.

Para el conjunto de datos original proporcionado, con un total de 8188, se considera que es un conjunto pequeño y este aspecto no afecta demasiado a la hora de decidir un modelo y otro. No obstante, si en trabajos futuros se procesa una cantidad superior, será un factor a considerar.

3. Conclusiones

Después de realizar un análisis exhaustivo de los datos proporcionados sobre los precios de las viviendas de Barcelona, se extraen las siguientes conclusiones:

1. Los precios de los alquileres tienen un sesgo positivo muy pronunciado, dejando una larga cola hacia los precios altos.
2. El tipo de vivienda más habitual es el tipo *flat*.
3. Los metros cuadrados del piso es el atributo que mayor coeficiente de correlación tiene con el precio de la vivienda, con un valor de 0.69, seguido por el precio por metro cuadrado y el número de baños, ambos atributos con un coeficiente de correlación de 0.58.
4. Hay un total de 8 atípicos en el conjunto de datos, concretamente en las filas: 6951, 1772, 2427, 2754, 4220, 4750, 7646, 7928.
5. Se han ajustado los siguientes modelos:
    - Regresión lineal simple
    - Lasso
    - Ridge
    - ElasticNet
    - K-nearest neighbor
    - Random Forest
    - Gradient Boosting
6. El modelo que ha proporcionado un menor error en el test ha sido Gradient Boosting, con un error de 20.53€.
7. El único modelo que ha salido beneficiado de la transformación logarítmica ha sido Gradient Boosting, con una reducción de 2.09€ respecto al error del modelo con la distribución original.
8. El modelo que más rápido se ha ajustado a los datos ha sido ElasticNet, con un tiempo de 37.4 ms.