Predicciones del precio de la vivienda a partir de datos del registro de Reino Unido

Domingo Parrales de la Cruz

Octubre 2024

Índice

1. Objetivos del proyecto **3**
1.1. Situación . . . 3
1.2. Complicación . . . 3
1.3. Objetivos . . . 3

2. Principales logros obtenidos **4**

3. Metodología **4**
3.1. Descripción de los objetivos . . . 4
3.2. Extracción, transformación y carga de datos . . . 4
3.2.1. Datos de Kaggle . . . 4
3.2.2. Datos censales . . . 13
3.2.3. Mapa de Reino Unido . . . 15
3.2.4. Posición de pueblos y ciudades . . . 15

4. Descripción de los modelos **17**
4.1. Modelos de regresión . . . 17
4.1.1. Variables del censo de Reino Unido . . . 17
4.1.2. Variables a partir de la información geográfica de Reino Unido . . . 18
4.1.3. Variables de los datos de Kaggle . . . 18
4.1.4. Selección de variables . . . 18
4.2. Series temporales . . . 18

5. Resultados principales **19**
5.1. Modelo de regresión . . . 19
5.2. Series temporales: predicción de la tendencia del precio medio de las transacciones en Reino Unido . . . 20

6. Recomendaciones de uso **23**

1. Objetivos del proyecto

1.1. Situación

Disponemos de datos sobre las ventas de viviendas en el Reino Unido desde enero de 1995 hasta abril de 2023. La mayor parte de la información se refiere a características geográficas, como el condado o el código postal correspondiente. También contamos con variables que describen características específicas de la vivienda, el tipo de contrato y la transacción realizada.

En la actualidad, estos datos podrían usarse para estimar el precio medio de una región en una fecha concreta. Sin embargo, aunque esto es de interés para la compra y venta de viviendas, puede no resultar suficiente. Por ejemplo, factores que facilitarían este tipo de transacciones incluirían la estimación del precio futuro de una vivienda o la predicción de las tendencias del mercado.

1.2. Complicación

Excepto el precio de venta de la vivienda, todas las variables son categóricas y la mayoría presentan una alta cardinalidad, es decir, un gran número de categorías únicas. Además, las variables que deberían proporcionar mayor precisión en la ubicación geográfica (más allá del pueblo o la ciudad) presentan una cantidad considerable de datos faltantes. Por ejemplo, hay un gran número de viviendas para las cuales no se especifica la calle.

Es necesario abordar este problema utilizando técnicas de ciencia de datos por varios motivos:

- En primer lugar, la construcción directa de un modelo es inviable debido a la alta cardinalidad de las variables categóricas. Es necesario encontrar otras formas de representar sus características.
- Además, resulta crucial identificar relaciones entre las variables para entender mejor el comportamiento del mercado.
- Finalmente, conocer los precios en función de la zona donde se encuentra la vivienda, junto con una predicción de su precio futuro, puede proporcionar una base sólida sin requerir un conocimiento previo y detallado de la zona.

1.3. Objetivos

- Predecir el precio de una vivienda según su ubicación y características básicas, así como de los precios medios en regiones más generales.
- Análisis de los principales factores que afectan al precio de la vivienda.
- Predicción de la evolución del precio medio de las transacciones.
- Predicción del tipo de vivienda que proporcionará un mayor retorno de inversión.

2. Principales logros obtenidos

Se ha logrado una alta capacidad de predicción de los precios medios mensuales por distrito, que también se aplica, aunque en menor medida, a viviendas concretas.

Además, se ha logrado capturar la tendencia real de los precios a lo largo del tiempo, lo que permite estimar las fechas con mayor retorno de inversión. Además, se ha analizado qué tipo de viviendas son de interés para su compra-venta, alquiler o construcción.

Finalmente, se han identificado las principales características que debe tener una vivienda para maximizar los beneficios.

3. Metodología

3.1. Descripción de los objetivos

En este proyecto afrontamos la predicción de precios con dos enfoques diferentes:

- Primero, hacemos un modelo de regresión, que podrá utilizarse tanto para estimar el precio futuro de viviendas concretas (especificando el tipo de vivienda, el tipo de contrato y la ciudad o pueblo al que pertenece) como para calcular el precio medio en regiones más amplias, especialmente a nivel de distrito.
- Por otro lado, se modeliza la evolución del precio de las transacciones basándose únicamente en su comportamiento pasado. No obstante, con este modelo se trabaja a nivel de Reino Unido completo, aunque sí que estudiaremos distintas categorías por separado, como pisos, viviendas de alquiler o viviendas de segunda mano. De esta forma, se podrá concretar qué tipo de vivienda dará un mayor retorno a la inversión.

3.2. Extracción, transformación y carga de datos

Trabajamos con cuatro fuentes de datos principalmente: los datos de Kaggle [1], datos censales de Reino Unido [2], un mapa de Reino Unido [3] e información geográfica sobre los pueblos y ciudades [4][5].

3.2.1. Datos de Kaggle

Estos datos contienen información sobre la venta de viviendas y están distribuidos según el año correspondiente. Esto resulta ventajoso, ya que, para evaluar la capacidad de predicción sobre casos futuros, se puede dividir el conjunto de datos en un conjunto de entrenamiento, que abarque desde enero de 1995 hasta diciembre de 2020, y un conjunto de prueba, que utilizará los datos desde enero de 2021 hasta abril de 2023.

Tomando una pequeña muestra de los datos y revisando la información proporcionada en la fuente original [6], podemos distinguir los siguientes tipos de variables en nuestro conjunto de datos:

- **Numéricas**: solo incluye el precio de venta de la vivienda. Se puede analizar su evolución y determinar si es necesario aplicar alguna transformación que mejore el rendimiento de los modelos.
- **Fechas**: la única variable de este tipo es la fecha de la transacción, de la cual se pueden extraer variables derivadas como el año, el mes y el día de la venta. Esto permitirá un análisis más detallado de las dependencias temporales en los modelos.
- **Categorías con baja cardinalidad**: las más relevantes son el tipo de vivienda, el tipo de tenencia de los inquilinos, la antigüedad de la vivienda y el tipo de transacción. Además, hay una variable que indica si fue necesario pagar tasas adicionales y otra que informa si la entrada fue modificada, añadida o eliminada de la base de datos.
- **Categorías con alta cardinalidad**: recogen la información espacial de los datos. Permiten un nivel de precisión hasta el número de la vivienda, aunque presentan una gran cantidad de datos faltantes. También incluyen un identificador único para cada transacción.

Además, siguiendo la descripción proporcionada por los datos [6], se pueden modificar los nombres de las categorías para que sean más comprensibles en el análisis. A continuación, se detalla el tratamiento aplicado a cada tipo de variable.

Categorías con baja cardinalidad

Dado que estas variables son las únicas que proporcionan información específica sobre la vivienda, es fundamental que se representen de la mejor manera posible. Para lograr esto, se tomará como referencia la variable objetivo y se analizará cómo influye cada categoría en ella.

Podemos comenzar con la variable *Old.New*, que indica si la vivienda es nueva o de segunda mano. Como se muestra en la Figura 1, la evolución del precio de venta para cada categoría presenta una clara diferenciación al inicio, pero, conforme avanza el tiempo, ambas siguen una progresión similar. Cabe destacar que, a pesar del ruido presente, se mantiene una tendencia clara.

Además, está el tipo de tenencia de los inquilinos, representado por la variable *Duration*. En la Figura 2, se puede observar que, a diferencia del caso anterior, la categoría 'Desconocido' introduce ruido en la variable. También se aprecia que esta categoría solo es relevante para datos pasados, por lo que es seguro eliminar estos casos, ya que no afectarán las predicciones futuras.

También se ha estudiado la variable *PPD.Category.Type*, que indica en qué casos se ha pagado un precio adicional por la propiedad, ya sea por tasas de embargo, herencia, etc. Como se muestra en la Figura 3, los casos en los que se pagan tasas adicionales no presentan un patrón claro hasta los últimos años, y la dispersión es considerablemente alta. Sin embargo, de la descripción de los datos [6] se extrae que la categoría 'Otro' de la variable *Property.Type* corresponde siempre a casos de venta con tasas adicionales, por lo que debemos comprobar también la relación que tiene con esa variable antes de descartar estos casos.

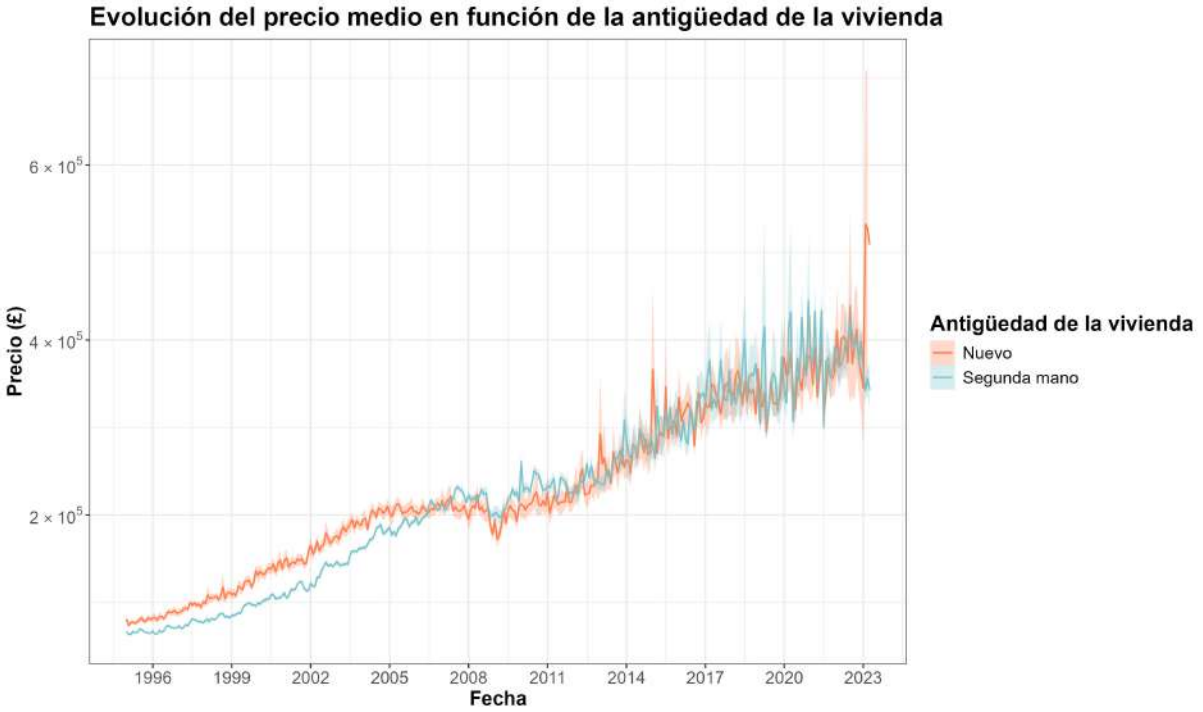

Figura 1: evolución del precio de la vivienda diferenciando entre el conjunto de propiedades nuevas y de segunda mano. Junto a la media se representa la dispersión del precio para esa fecha como un área semitransparente. En ambas se tiene una evolución similar.

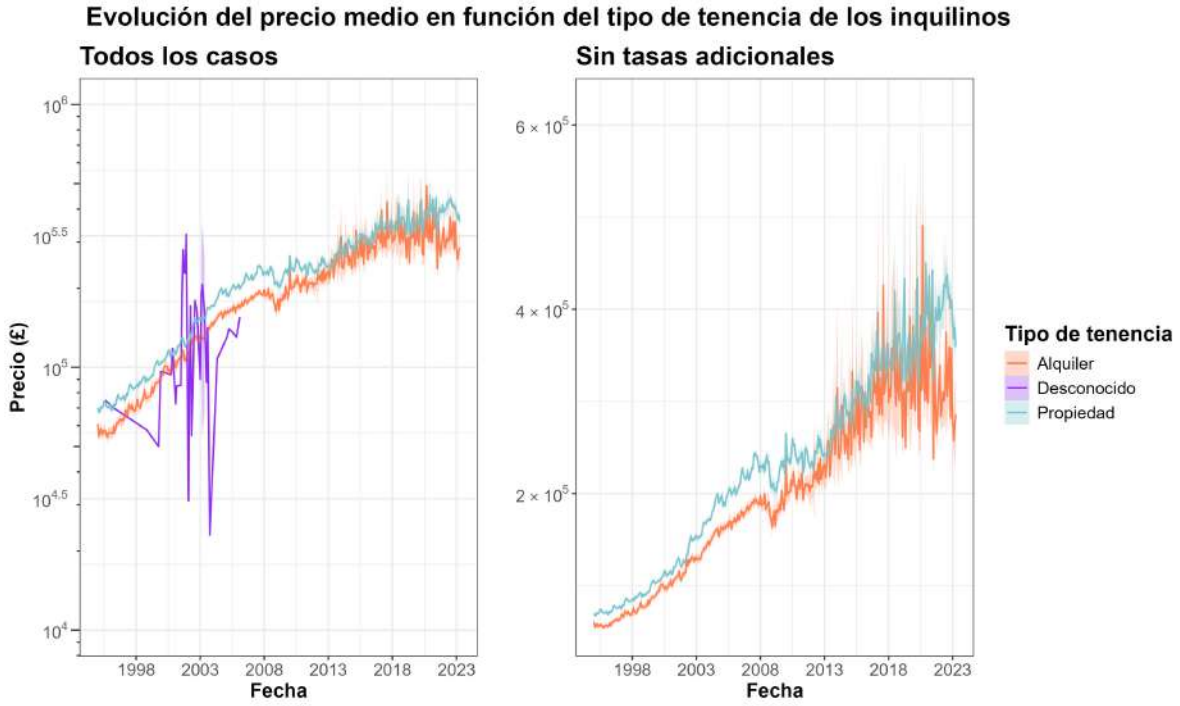

Figura 2: evolución del precio de la vivienda según el tipo de tenencia de los inquilinos. Podemos ver que la categoría 'Desconocido' solo ocurre en datos pasados. Se representa, junto a la línea del valor medio, la dispersión para esa fecha mediante el área semitransparente que la acompaña.

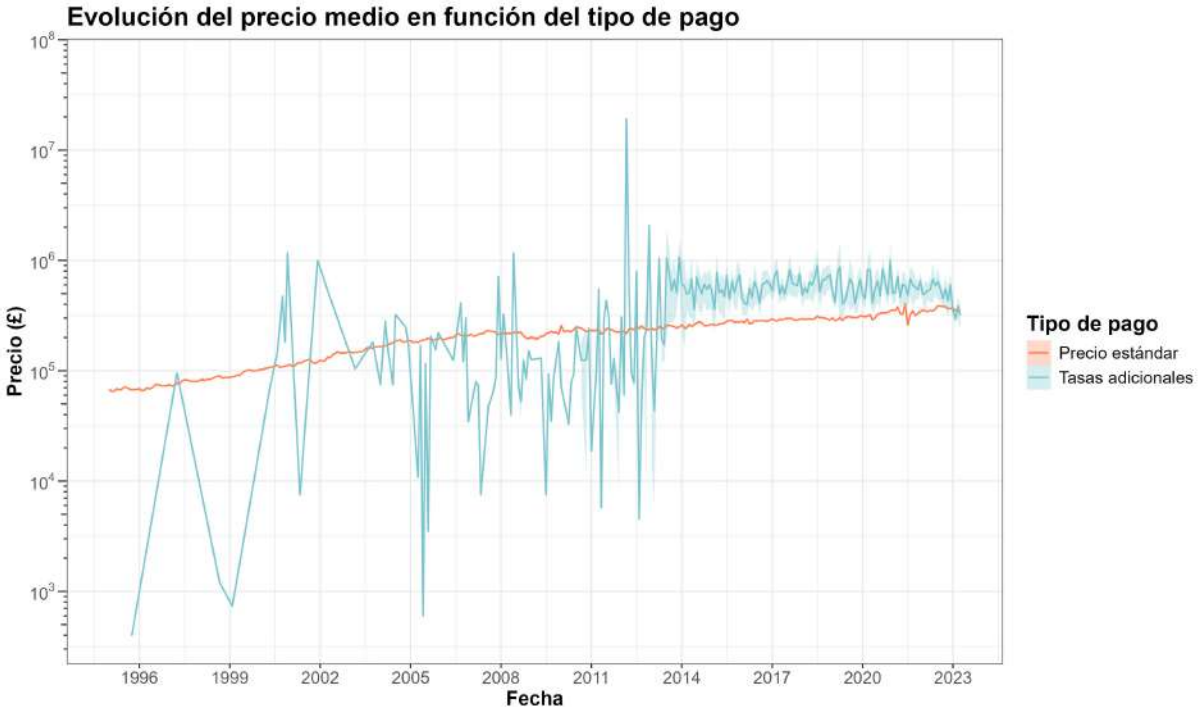

Figura 3: evolución del precio de la vivienda en función de si se pagaron o no tasas adicionales en la transacción. Junto al valor medio, se muestra una zona sombreada que representa la dispersión para esa fecha. Se aprecia que los casos en los que se pagó un precio adicional presentan una dispersión muy elevada.

Por último, está la variable *Property.Type*, que nos indica el tipo de propiedad. Como se muestra en la Figura 4, la variable 'Otro' presenta un problema idéntico al anterior. No obstante, se puede ver que son una minoría al fijarnos en la Figura 5 (el tamaño de los cuadrados representa el número de casos que hay para cada pareja). En consecuencia, se puede descartar la categoría para conservar la estabilidad del modelo, como se observa a la derecha en la Figura 4.

Categorías con alta cardinalidad

Especificar la ubicación geográfica de una vivienda mediante nombres tiene la ventaja de que se puede comparar con otras bases de datos con relativa facilidad. Sin embargo, pueden existir diferencias, como en el uso de puntuación, la ausencia de *stopwords* (como 'the' u 'of') o incluso el uso de abreviaturas ('St.' en lugar de 'Saint').

Esto plantea un preprocesamiento que puede volverse inabarcable dependiendo del nivel de precisión o granularidad deseado. Para tomar una decisión, se ha calculado el número de categorías únicas que presenta cada variable geográfica, como se muestra en el Cuadro 1. El número de categorías aumenta progresivamente hasta llegar a la localidad, donde se observa un incremento drástico. Considerando esto, se ha decidido establecer como nivel máximo de precisión los pueblos y ciudades, aunque el análisis se centrará principalmente en los distritos.

Además, es importante señalar que los resultados de el Cuadro 1 no consideran el número de

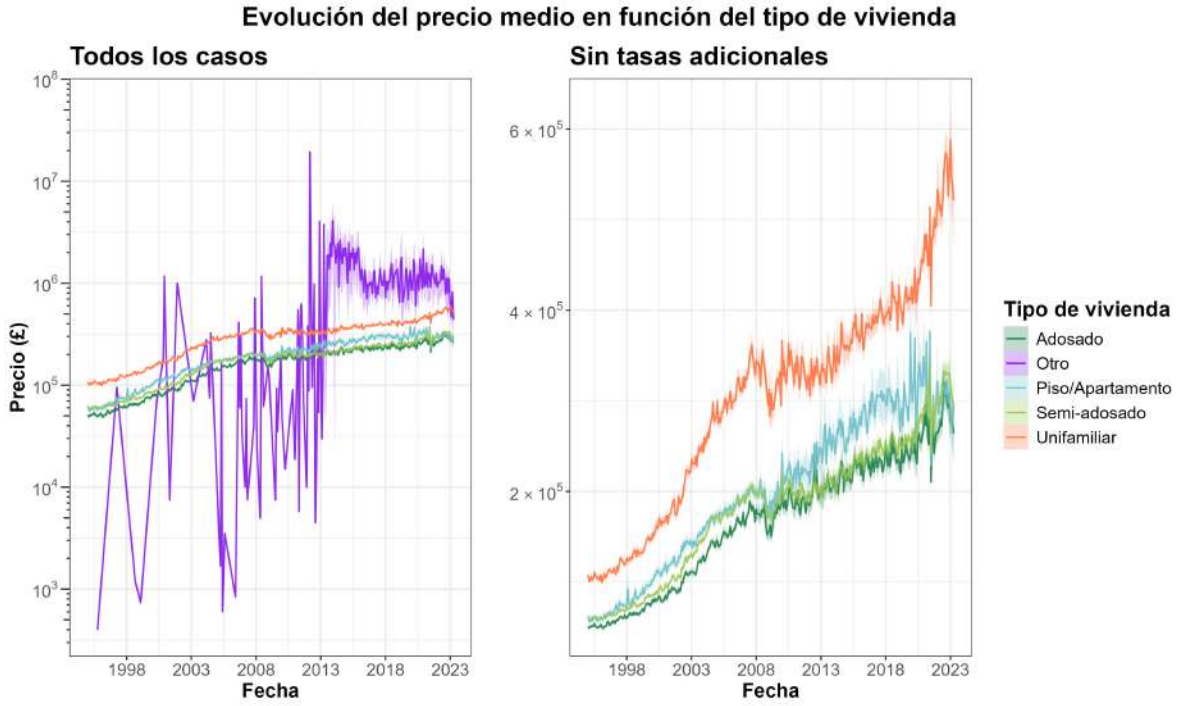

Figura 4: evolución del precio de la vivienda segregado según el tipo de propiedad. En la imagen izquierda vemos que la categoría 'Otro' introduce una gran dispersión, interfiriendo en la caracterización de la tendencia de las otras categorías, que se muestran en la imagen derecha.

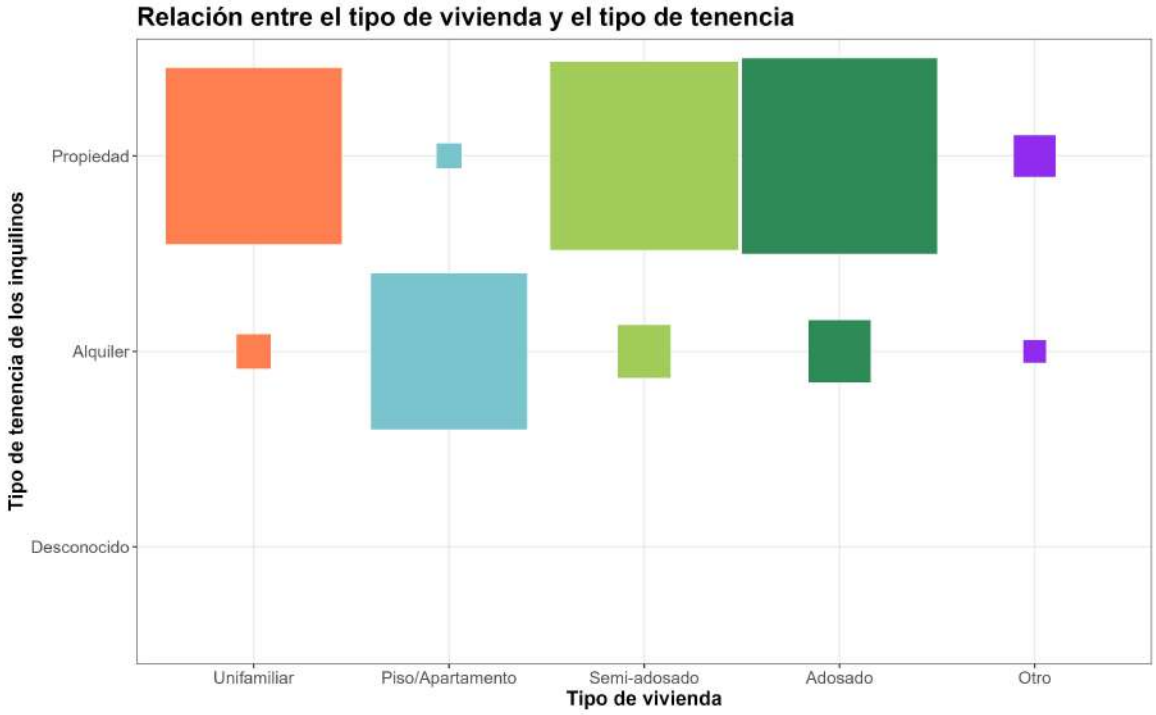

Figura 5: representación del número de casos para cada par *Property.Type* - *Duration*. Esta cualidad queda representada por el tamaño del cuadrado. Podemos ver que la categoría 'Otro' de *Property.Type* y 'Desconocido' de *Duration* son una minoría en los datos.

Variable	Número de categorías únicas
Condado	132
Distrito	467
Pueblo/Ciudad	1172
Barrio	23862
Calle	326400

Cuadro 1: número de categorías únicas de algunas de las variables geográficas disponibles. Las ausentes hacen referencia al código postal y el número de la vivienda, pero tienen una cantidad de datos faltantes excesiva.

combinaciones posibles, derivadas de la estructura jerárquica de estas variables[1]. En concreto, si se considera la multiplicidad de combinaciones en los datos de Condado - Distrito - Pueblo o Ciudad, se obtienen 7563 casos distintos.

En definitiva, se descartará cualquier variable que proporcione una precisión mayor que la de pueblos o ciudades. Esto también evita el uso de variables con datos faltantes. En la Figura 6 se muestran las relaciones entre los datos faltantes de las variables. A partir de esta figura, podemos deducir lo siguiente:

- A mayor precisión geográfica, más datos faltantes.
- La variable *SAON* es la principal fuente de datos faltantes. Esto es comprensible, ya que solo se utiliza cuando es necesario añadir información adicional al número de la casa (*PAON*), como en el caso de indicar el número de piso.
- La ausencia de datos en *Locality* sugiere que hay propiedades sin un barrio asignado. Esto es posible, especialmente en áreas rurales o aisladas, aunque no debería ser tan frecuente.
- Existen casos en los que faltan exclusivamente el código postal o el nombre de la calle. A diferencia del caso anterior, aquí realmente falta información esencial que será difícil de imputar.

Variable objetivo: precio de venta

Al construir un modelo de regresión, uno de los factores más importantes es considerar la normalidad de la variable objetivo, ya que muchos modelos suponen esta característica. Para comprobar si este es nuestro caso, se ha representado a la izquierda de la Figura 7 la distribución de los valores del precio de las transacciones.

Es evidente que la distribución de precios está fuertemente sesgada a la izquierda, lo que indica que la mayoría de los precios están concentrados en valores bajos, mientras que las propiedades con precios de venta más altos son una minoría.

[1]Un condado contiene distritos. Un distrito contiene pueblos y ciudades. Los pueblos y ciudades contienen barrios, y los barrios contienen calles.

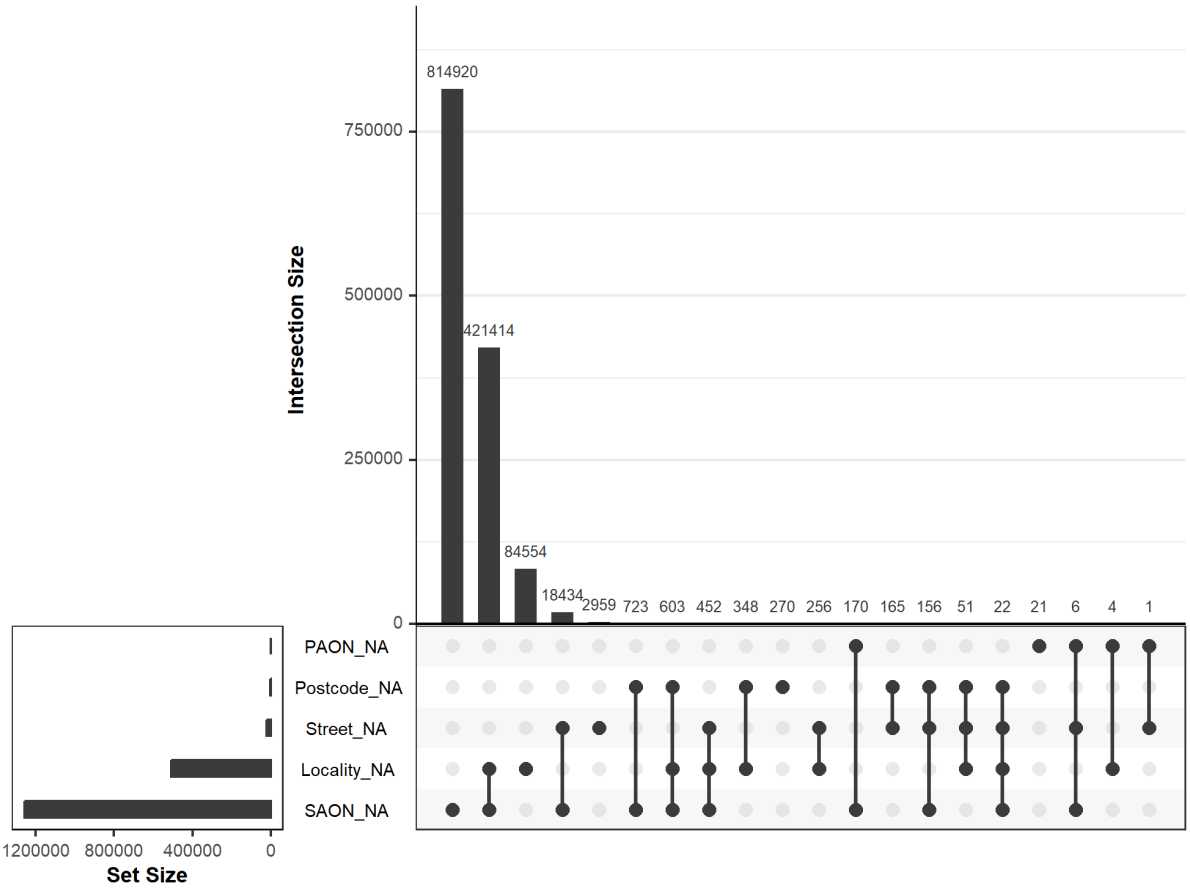

Figura 6: distribución de datos faltantes entre las variables de los datos de Kaggle creado con la librería *naniar* de *R*. Con las barras horizontales de la izquierda, podemos ver la cantidad de datos faltantes en las variables afectadas. Con los puntos en la zona inferior derecha se representan las conexiones entre los datos faltantes. Las barras verticales representan el número de casos en los que se cada cada relación marcada con los puntos.

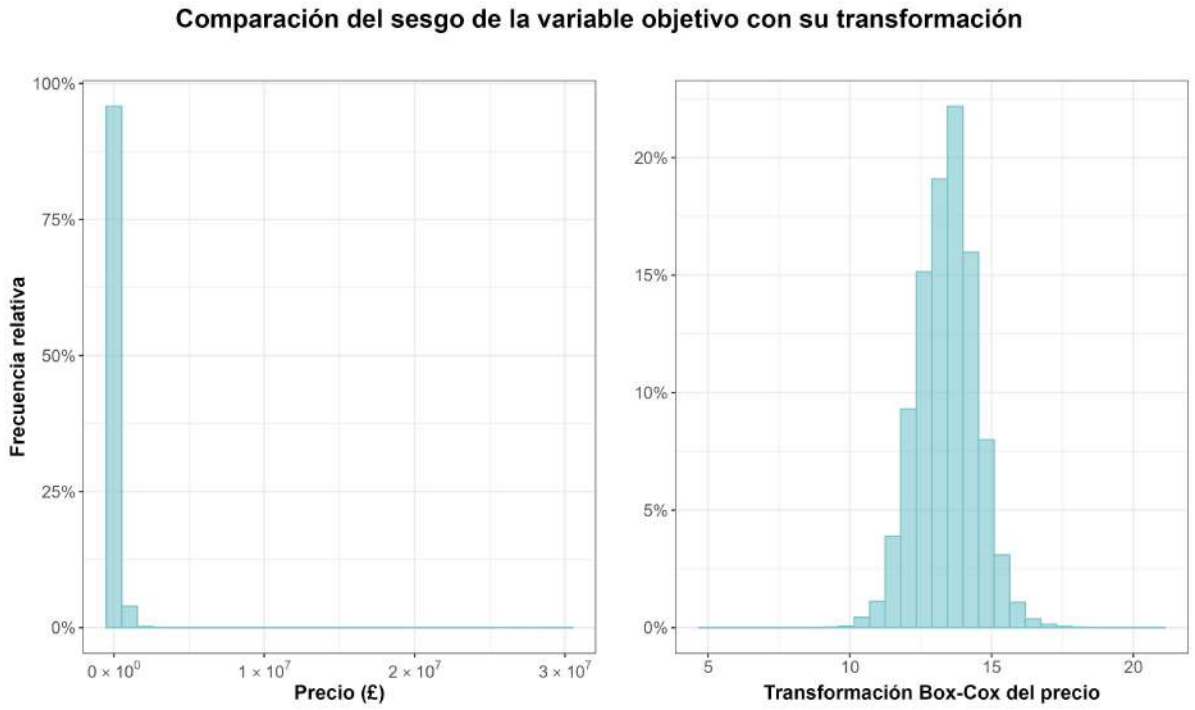

Figura 7: A la izquierda representamos el número de casos en la horquilla de precios que representa el ancho de cada barra. A la derecha, se tiene una representación análoga, pero con la variable objetivo transformada (mediante una transformación Box-Cox).

Una forma de corregir este sesgo es mediante la transformación de la variable. En particular, se ha optado por aplicar la transformación Box-Cox a los datos y, como se muestra a la derecha en la Figura 7, la distribución resultante se ajusta de manera más cercana a una distribución gaussiana.

Outliers

A partir del análisis de normalidad de la variable objetivo, se ha identificado que los precios más altos representan una minoría en el conjunto de datos. Sin embargo, se puede extraer más información comparando los *outliers* de los datos históricos (desde enero de 1995 hasta abril de 2023) con los de los más recientes (desde el 2020 hasta abril del 2023).

Para ello, se analizan las Figuras 8, 9 y 10, en las que las cajas representan donde se acumula el 50 % de los datos y la línea en su interior es la mediana, que podemos tomar de referencia para ver entorno a qué valores se acumulan los precios. Además, los puntos representan los casos que no se asemejan lo suficiente a los de la caja. De ellas, se extraen varias características comunes a las tres:

- La mediana de los datos más recientes es consistentemente mayor que la mediana de los precios históricos. Esto indica un crecimiento económico en el sector inmobiliario del Reino Unido.
- Para los datos actuales, los puntos de los precios más bajos se alejan menos de la mediana que los de precios más altos. Esto implica que, en general, los casos excepcionales suelen

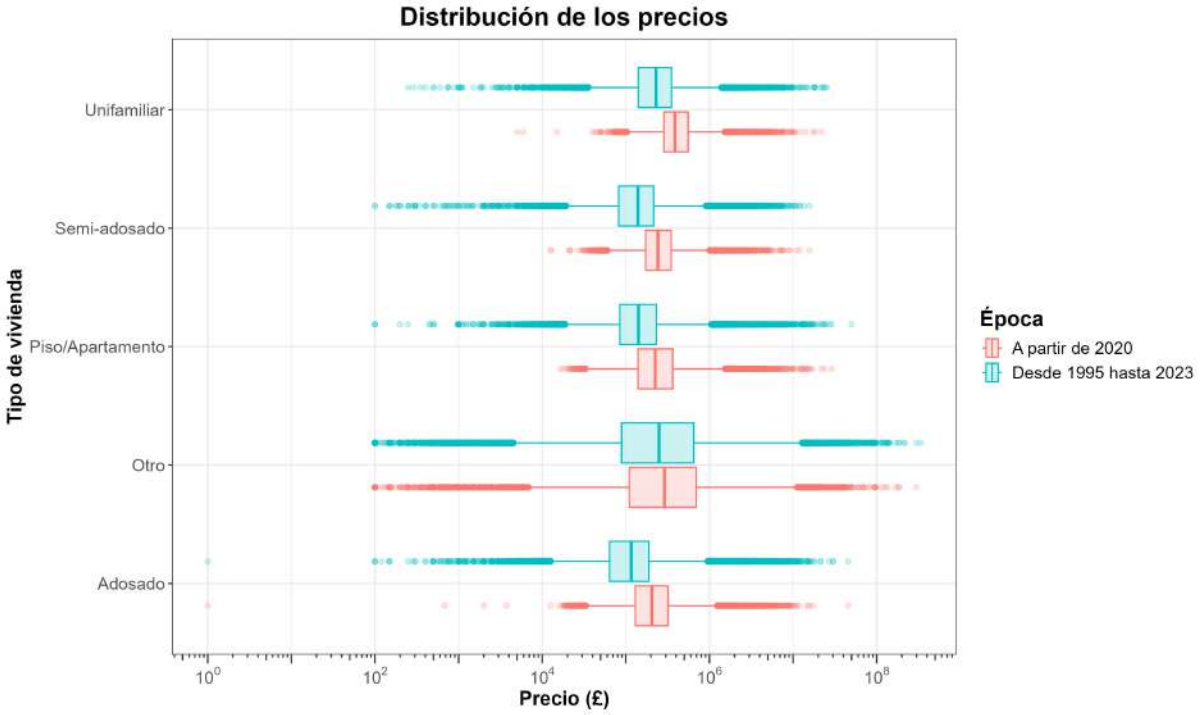

Figura 8: boxplot del precio de venta de las propiedades, donde se han dividido los datos en categorías. Se representan conjuntamente los datos más recientes (desde enero de 2020 hasta abril de 2023) junto a los históricos (desde enero de 1995 hasta abril de 2023). Se utiliza escala logarítmica en el precio para facilitar la visualización. Se observa que, en la mayoría de casos, los precios actuales superan a los históricos.

dar mayores beneficios.

- Los puntos correspondientes a los precios más elevados en los datos históricos, suelen ser mayores a los puntos de precios más elevados de los precios actuales. Esto se aprecia especialmente en la Figura 10, por ejemplo, para las viviendas con contratos de alquiler, en la que los puntos más a la derecha se dan para los precios históricos.

Además, se puede extraer información relevante sobre la evolución del mercado:

- En la Figura 8 se observa que, sin tener en cuenta la categoría *Otro*, los precios más elevados suelen involucrar viviendas unifamiliares. No obstante, si tenemos en cuenta los puntos, estos precios son comparables al de algunos pisos/apartamentos y viviendas adosadas, aunque en menor medida.
- De la Figura 9, se destaca que los precios de las viviendas nuevas están más concentrados entorno a la mediana que los de las viviendas de segunda mano. Esto es especialmente evidente en los precios de 2020.
- Finalmente, la Figura 10 muestra que, en general, los contratos de propiedad son más caros que los de alquiler, aunque la diferencia no es tan grande.

Aunque los puntos corresponden a casos excepcionales, por lo que podrían afectar negativamente a las predicciones. Por este motivo, se han eliminado de los datos usados para estudiar la evolución del precio de las transacciones, aunque se han mantenido en los datos para el modelo de

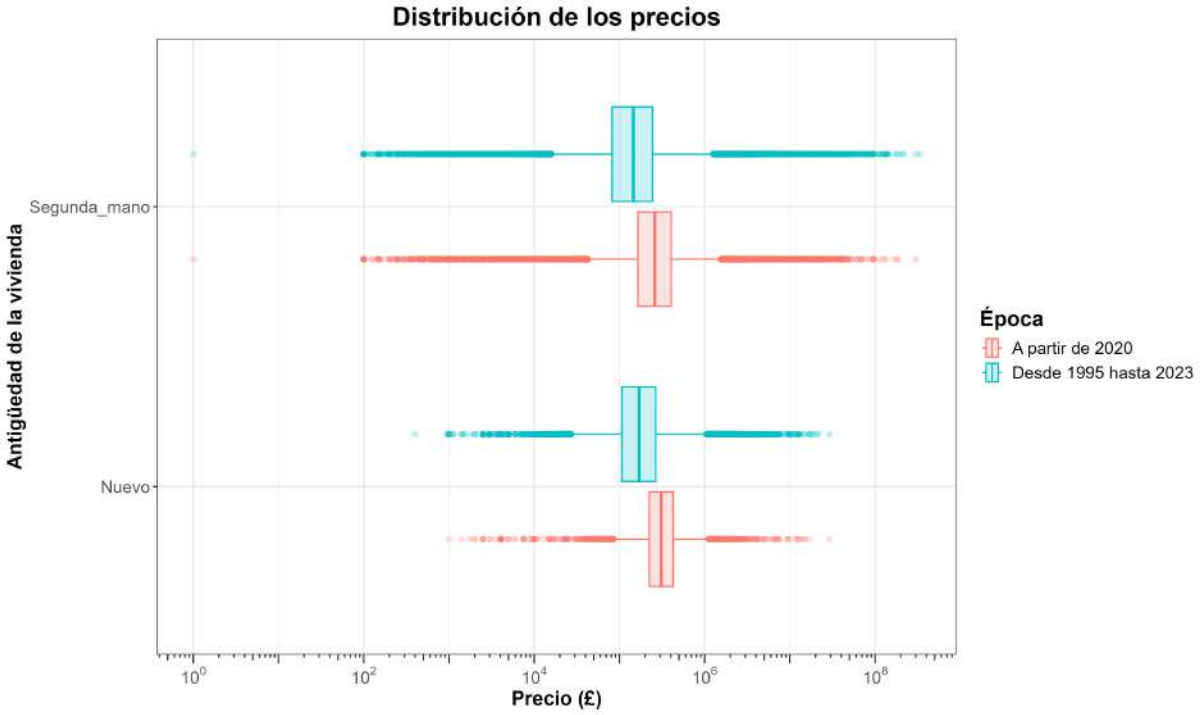

Figura 9: boxplot del precio de venta de las propiedades, donde se han dividido los datos en categorías. Se representan conjuntamente los datos más recientes (desde enero de 2020 hasta abril de 2023) junto a los históricos (desde enero de 1995 hasta abril de 2023). Se utiliza escala logarítmica en el precio para facilitar la visualización. Se observa que, en la mayoría de casos, los precios actuales superan a los históricos.

regresión. Esto se debe a que, para este último, se cuentan con más variables además del precio. En concreto, se añaden variables con las que clasificar cada registro en función de si excede los precios por encima de lo esperado o por debajo.

3.2.2. Datos censales

Esta es la principal fuente de enriquecimiento de los datos originales. Cuenta con varias categorías relevantes como la densidad de trabajo, la cantidad de casas o el número de estudiantes, todas ellas referente a los distritos de las regiones de Inglaterra y Gales, que son las únicas que se tienen en los datos originales.

La principal ventaja de estos datos es que proporcionan una representación numérica de los distritos, lo que permite representarlos sin necesidad de codificarlos como una variable categórica.

Sin embargo, la incorporación de estos datos no es completamente directa, ya que los nombres de algunos distritos no coinciden con los de los datos de Kaggle. En el Cuadro 2 se muestran algunos de los cambios realizados.

Otro factor a tener en cuenta es que los datos extraídos corresponden a un año concreto. Se ha optado por utilizar los más recientes, lo que, en principio, favorecerá las predicciones futuras y podría afectar negativamente a las predicciones pasadas. Además, para los distritos que desapa-

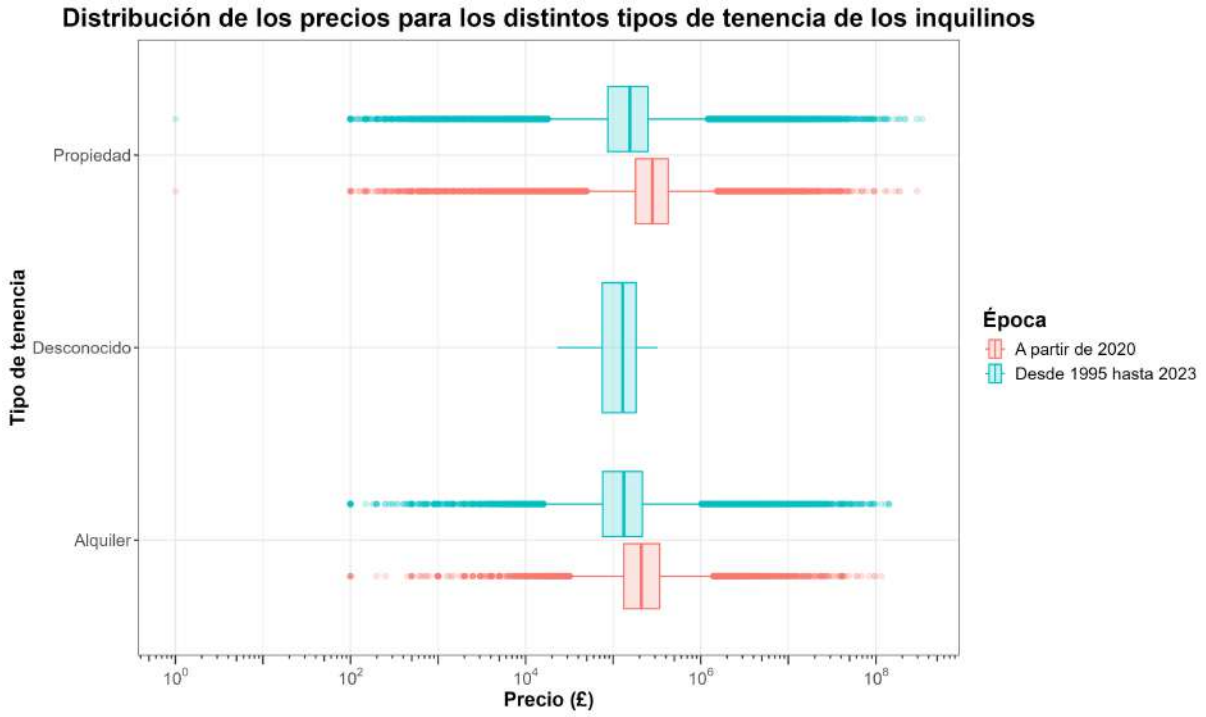

Figura 10: boxplot del precio de venta de las propiedades, donde se han dividido los datos en categorías. Se representan conjuntamente los datos más recientes (desde enero de 2020 hasta abril de 2023) junto a los históricos (desde enero de 1995 hasta abril de 2023). Se utiliza escala logarítmica en el precio para facilitar la visualización. Se observa que, en la mayoría de casos, los precios actuales superan a los históricos.

Distritos de los datos censales	
Original	**Corrección**
County Durham	Durham
St. Helens	St Helens
Bristol, City of	Bristol

Distritos del mapa de Reino Unido	
Original	**Corrección**
Rhondda, Cynon, Taff	Rhondda Cynon Taff
St. Helens	St Helens
Telford and Wrekin	Wrekin

Cuadro 2: ejemplos de correcciones realizadas a los nombres de distritos en el conjuntos de datos del censo de Reino Unido y los de su mapa. El objetivo de dichos cambios es que coincidan con los del conjunto de datos original. Destacan la eliminación de puntuación y de términos como 'County' o 'City'.

recieron antes del registro, no habrá ningún valor asociado, por lo que se han imputado estos valores faltantes utilizando la media de la variable correspondiente.

3.2.3. Mapa de Reino Unido

Se ha utilizado una representación en polígonos de los distritos actuales en el Reino Unido para generar visualizaciones geográficas de los datos.

La principal dificultad en el uso de estos datos es similar a la que presentan los datos censales: es necesario corregir algunos nombres de distritos y, dependiendo de lo reciente que sea el mapa, los distritos antiguos de nuestros datos no estarán representados. Algunos ejemplos de las correcciones realizadas en este conjunto de datos se encuentran a la derecha en el Cuadro 2.

Utilizando estos datos, se ha elaborado en la Figura 11 una representación espacial de la media de precios por distrito. A partir de esta representación, se pueden observar tres características fundamentales:

- En primer lugar, se confirma nuevamente que, en general, los precios de las transacciones han aumentado en los últimos años, ya que se tienen más colores cercanos al amarillo.
- En segundo lugar, se observa que, a medida que nos acercamos a Londres, los precios aumentan rápidamente. Esto es visible tanto en los precios recientes como al considerar todos los precios registrados. Este patrón puede incluirse como variable en el modelo para capturar parte de la relación espacial entre los distritos.
- Finalmente, se aprecia que, en general, hay agrupaciones de distritos con color similar. Por lo tanto, parece razonable introducir una variable adicional que agrupe los pueblos en función de su proximidad espacial.

3.2.4. Posición de pueblos y ciudades

Hasta ahora, se ha limitado la precisión geográfica a los distritos para reducir las correcciones necesarias al adjuntar otros conjuntos de datos. Sin embargo, si se utilizan los nombres presentes en nuestros datos, no será necesario este tratamiento.

Teniendo esto en cuenta, se ha decidido emplear los nombres de pueblos y ciudades para extraer su posición geográfica (latitud y longitud). Para ello, en primer lugar, se ha utilizado la API de OpenCage [4], y posteriormente se revisaron los casos con valores atípicos o que quedaron sin asignar mediante la API de Nominatim [5][2].

Para hacer las peticiones, inicialmente se proporcionaron direcciones a la API de OpenCage con la estructura 'Condado, Distrito, Pueblo/Ciudad'. Sin embargo, algunas coincidían con lugares

[2]La API de OpenCage tiene un límite de uso diario generoso en su versión de prueba, mientras que Nominatim es Open Source, por lo que se redujo el ritmo de peticiones al usarla. Existen otras opciones como la API de Google [7], que permite un gran número de peticiones y podría resultar más atractivo para proyectos a gran escala.

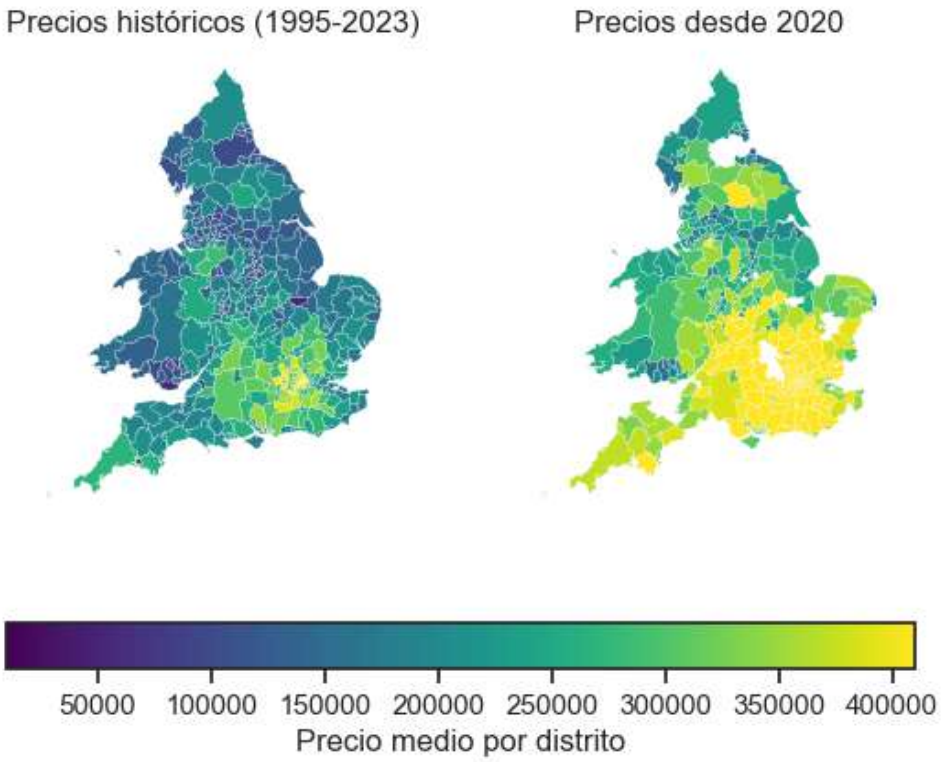

Figura 11: mapa de Reino Unido de las regiones de Inglaterra y Gales. Se muestra, según la escala de color inferior, la distribución geográfica del precio medio de venta de la vivienda por distrito. Para la imagen izquierda, usamos todos los casos disponibles, mientras que en la derecha solo los que se registran tras 2020.

en Australia. Posteriormente, con la API de Nominatim, se empleó la estructura 'Pueblo/Ciudad, UK', dando mayor importancia al nombre del pueblo o ciudad y especificando que pertenecen al Reino Unido.

4. Descripción de los modelos

Tanto para los modelos de regresión como para los de series temporales, el objetivo es obtener un modelo que pueda predecir precios futuros. Por tanto, como los datos utilizados abarcan desde enero de 1995 hasta abril de 2023, hemos decidido mantener los datos a partir de ciertos años como conjunto de test. En concreto:

- Para los modelos de regresión tomamos como conjunto de entrenamiento el rango entre enero de 1995 y diciembre de 2020. Por tanto, se tienen como conjunto de test los datos entre enero de 2021 y abril de 2023. Además, se toma una muestra aleatoria del 5 % de los datos de test como conjunto de validación.
- Para los modelos de series temporales se toma el mismo rango para realizar el estudio inicial. Sin embargo, para el ajuste de hiperparámetros, se pasa a tener tres conjuntos: un conjunto de entrenamiento, con datos entre enero de 1995 y diciembre de 2005; un conjunto de validación, con datos entre enero de 2006 y diciembre de 2015; y un conjunto de test, con datos entre enero de 2016 y abril de 2023. Al mantener el conjunto de test al margen del entrenamiento, es posible verificar la capacidad de predicción del modelo.

También, se contaba inicialmente con aproximadamente 30 millones de registros, restringidos a las regiones de Inglaterra y Gales. Sin embargo, en base a la información extraída del análisis inicial, se han filtrado los casos para centrarse en:

- Ventas correspondientes a viviendas adosadas, semi-adosadas, unifamiliares o pisos/apartamentos.
- Contratos de venta o alquiler.
- Viviendas nuevas o de segunda mano.
- Casos que no son considerados outliers.

4.1. Modelos de regresión

Para mejorar la capacidad predictiva de los modelos de regresión, se han añadido múltiples variables a partir de los conjuntos de datos disponibles. Para cada uno se ha enfocado la generación de variables en ciertos ámbitos concretos.

4.1.1. Variables del censo de Reino Unido

Del censo se ha extraído información sobre los distritos de diversa índole, que puede agruparse en dos categorías:

- **Información socio-demográfica**: número total de personas por distrito y cantidad de ellas que son trabajadoras o estudiantes, la densidad de empleo por habitante y la densidad de población (por hectárea).
- **Información de las viviendas**: número total de viviendas en cada distrito y el porcentaje de estas en cada nivel de eficiencia energética (nivel A, nivel B, etc.).

Todas las variables seleccionadas son numéricas, lo que permite representar los distritos en función de algunas de sus características clave. Para su selección, se consideraron los factores que, en principio, se cree que pueden influir en el precio de la vivienda.

4.1.2. Variables a partir de la información geográfica de Reino Unido

De manera directa, no se disponen de nuevas variables para añadir a los datos disponibles. Sin embargo, es posible crear algunas que aporten información adicional. En concreto, se ha añadido:

- **Información sobre los distritos**: si se encuentra en zona costera, si es una isla, su distancia a Londres.
- **Información sobre los pueblos y ciudades**: se han hecho grupos de pueblos y ciudades que están especialmente cerca entre sí, ya que se podrían influir los unos a los otros.

4.1.3. Variables de los datos de Kaggle

Además de la información ya disponible, las variables creadas principalmente tratan sobre información económica de los distritos. Por ejemplo, se ha analizado la diferencia de precio respecto al año anterior para cada distrito.

4.1.4. Selección de variables

Para mejorar el rendimiento de los modelos, es necesario hacer un estudio sobre la relación entre las variables, pero también es posible extraer información adicional sobre el mercado. De esta forma, se ha encontrado que:

- La distancia a Londres representa relativamente bien el descenso del precio de las transacciones.
- El valor D de eficiencia energética, conlleva, en general, un descenso del precio de la vivienda, mientras que el valor C supone un aumento. Por tanto, podría resultar interesante reformar una vivienda para que pasase del nivel D al C antes de venderla.
- El factor socio-demográfico que mejor representa los datos es la densidad de población de un distrito.

4.2. Series temporales

Se ha desarrollado un modelo con el que estudiar el comportamiento en el tiempo del precio de las transacciones. Para entrenarlo, únicamente se usa el comportamiento pasado de los precios

Error en la predicción de categorías concretas	
Categoría	**MSE ()**
Contrato de alquiler	56200
Contrato de propiedad	47900
Vivienda de segunda mano	41700
Vivienda nueva	97300
Vivienda adosada	57600
Piso/Apartamento	59100
Vivienda semi-adosada	43300
Vivienda unifamiliar	93200

Cuadro 3: medidas del error (MSE) de estimación de cada categoría por separado. Se aprecia que las debilidades principales del modelo son las viviendas nuevas y las unifamiliares.

y, a partir de él, se hace una estimación de la tendencia que presentará (crecimiento o decrecimiento a rasgos generales) y su comportamiento estacional (cambios que se producen de forma reiterada en el tiempo).

Uno de los factores principales al trabajar con este tipo de modelos es la elección de la precisión temporal con la que queremos trabajar (diaria, mensual o anual). En este proyecto, se ha optado por la precisión mensual, ya que es capaz de capturar la mayoría de tendencias diarias, mientras que también tiene en cuenta las tendencias a largo plazo.

Además, se han construido modelos con las mismas características, pero entrenados en conjuntos de datos concretos, por ejemplo, en los registros de viviendas de segunda mano. De esta forma, es posible distinguir qué sector es más llamativo para la inversión y podemos identificar las debilidades del modelo. Sobre esto último, en el Cuadro 3 se aprecia que las categorías con mayor error y, en consecuencia, los menos precisos, son las viviendas nuevas y las viviendas unifamiliares.

5. Resultados principales

5.1. Modelo de regresión

El modelo de regresión desarrollado presenta un gran rendimiento en la predicción de precios medios de distritos. Estas estimaciones pueden realizarse con distintas precisiones temporales (precios medios por distritos diarios, mensuales o anuales), pero se considera que la más adecuada es la mensual, ya que la diaria tiene un rendimiento bastante menor y la anual no presenta una mejora significativa a cambio de la pérdida de información por pasar a esta escala.

Además, podemos usar el mismo modelo para predecir los precios de viviendas concretas, aunque, como se muestra en el Cuadro 4, el rendimiento en este contexto es considerablemente inferior. Se piensa que la razón de los resultados tan dispares radica en que, para los distritos, se disponía de información adicional mediante los datos censales, mientras que para las viviendas individuales solo se contaba con características básicas. Sería interesante explorar la posibilidad

Modelo de regresión evaluado en test		
	R^2 **medio**	MSE **medio**
Viviendas concretas	0,56	0,32
Pueblos y ciudades (diario)	0,57	0,14
Distritos (diario)	0,67	0,14
Condados (diario)	0,53	0,27
Pueblos y ciudades (mensual)	0,55	0,32
Distritos (mensual)	0,83	0,20
Condados (mensual)	0,71	0,07
Pueblos y ciudades (anual)	0,56	0,19
Distritos (anual)	0,86	0,04
Condados (anual)	0,74	0,06

Cuadro 4: resultados de evaluar el modelo de regresión sobre el conjunto de test completo, es decir, los datos entre inicios de 2021 y abril de 2023. El primer caso es la evaluación del acierto en la predicción del precio de viviendas concretas. Los siguientes son el acierto en la predicción del precio medio para un cierto grado de precisión geográfica y temporal.

de incluir información más detallada, como los metros cuadrados de la vivienda y el número de habitaciones, ya que esto podría mejorar significativamente el rendimiento del modelo en la predicción de viviendas específicas.

En resumen, este modelo no solo permite un análisis actual de los precios medios por región, sino que también ofrece una predicción de cómo evolucionarán en los próximos años, lo que representa una ventaja considerable a nivel de inversión en el sector inmobiliario.

5.2. Series temporales: predicción de la tendencia del precio medio de las transacciones en Reino Unido

Para complementar las predicciones de precios realizadas con el modelo anterior, se han estimado las tendencias del precio medio de las transacciones en Reino Unido. Se ha encontrado que con un modelo autorregresivo puede representarse con un error de aproximadamente el 10 % del valor estimado, lo que permite realizar predicciones relativamente precisas.

Después, se han entrenado modelo con los parámetros en datos más específicos, como viviendas con contrato de alquiler, de segunda mano o unifamiliares. De esta forma, se ha identificado que las categorías peor representadas por el modelo son las viviendas nuevas y las unifamiliares, ambas duplicando el error que presentan las otras categorías, como se muestra en el Cuadro 3.

Además, se han aprovechado estas predicciones para hacer una comparación entre la evolución de los precios de las distintas categorías, lo que ayudará a elegir en qué sector invertir. Esto queda representado en la parte derecha de las Figuras 12, 13 y 14, de las que se extrae la siguiente información:

- En todas las categorías se observa un crecimiento en el precio, aunque este difiere en términos de velocidad y estabilidad a lo largo del tiempo. En principio, las categorías más interesantes serán aquellas con un crecimiento rápido o que presenten inestabilidades, lo

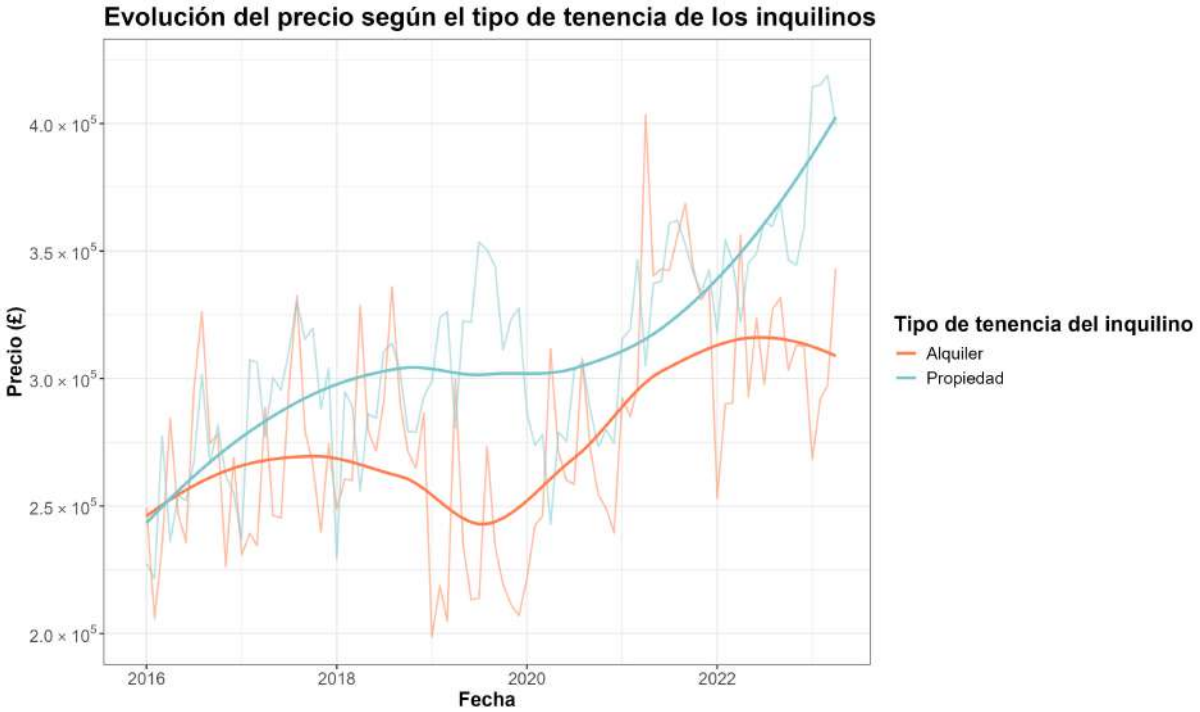

Figura 12: evolución del precio en función del tipo de tenencia de los inquilinos. La línea principal indica la tendencia que siguen los datos semi-transparentes que se aprecian entorno a la curva.

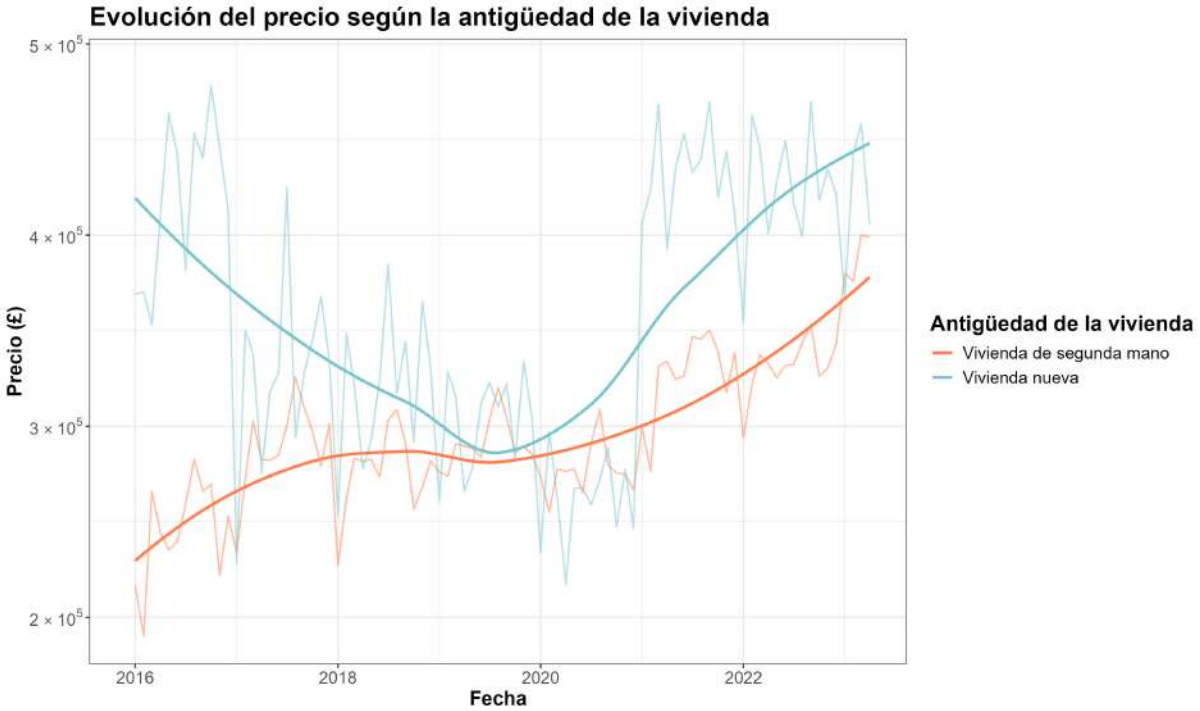

Figura 13: evolución del precio en función de la antigüedad de la vivienda. La línea principal indica la tendencia que siguen los datos semi-transparentes que se aprecian entorno a la curva.

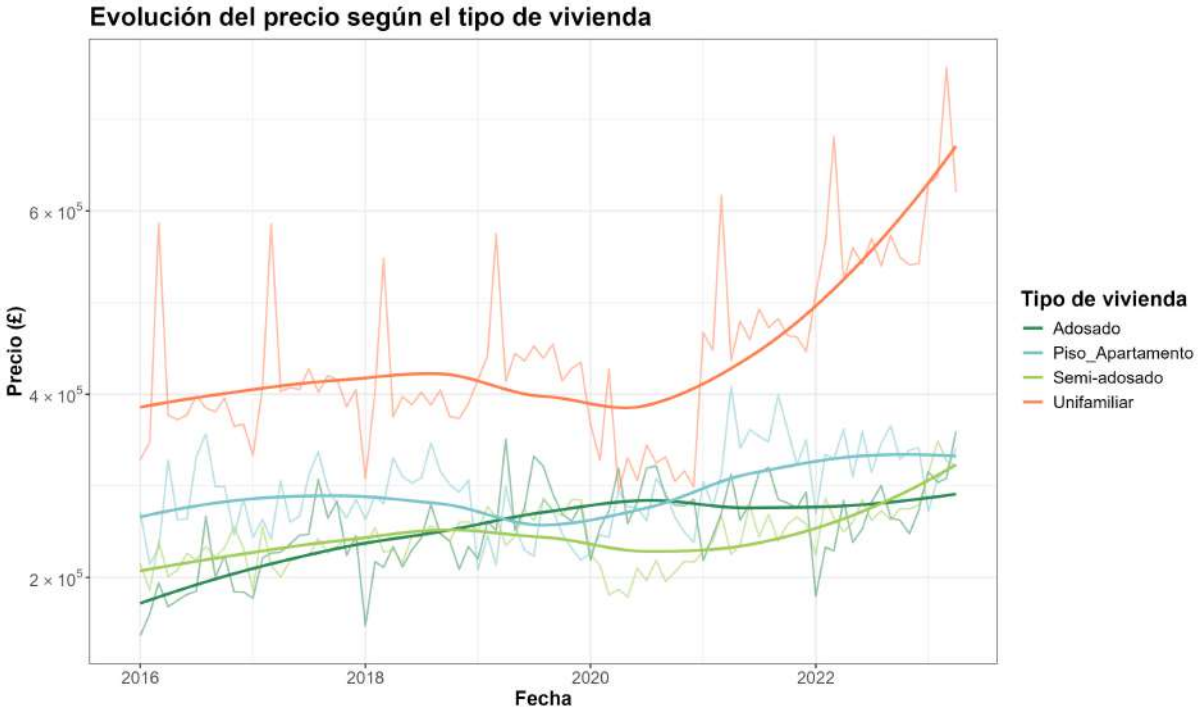

Figura 14: evolución del precio en función del tipo de vivienda. La línea principal indica la tendencia que siguen los datos semi-transparentes que se aprecian entorno a la curva.

que permitiría invertir en momentos de baja del mercado y vender en momentos de alza. No obstante, la estabilidad en los precios ayuda en la estimación del retorno de inversión de la construcción de viviendas.

- En la Figura 12 se observa que la evolución de los precios de los contratos de propiedad tiene un crecimiento más estable que la de contratos de alquiler. Usando la información disponible, se podrían hacer contratos de alquiler que se renueven en los momentos de alza del mercado y evitar momentos en los que el precio baja drásticamente.

- En la Figura 13 lo más llamativo es el repentino crecimiento de las viviendas nuevas. Esto podría deberse al gran error de estimación que presenta esta categoría concreta, aunque cabe destacar que ocurre muy cerca de la pandemia de COVID-19 y del Brexit. Por otro lado, se observa un crecimiento relativamente rápido del precio de las transacciones de viviendas de segunda mano.

- En la Figura 14, se observa que las propiedades unifamiliares también presentan un incremento similar al de las viviendas nuevas (de la misma magnitud, aunque parece menor por la escala del eje vertical). De las otras tres opciones, las viviendas semi-adosadas son las que tienen un crecimiento más marcado, aunque las adosadas presentan una estabilidad muy interesante para el sector de la construcción.

En resumen, si se busca invertir en viviendas para venderlas, las más interesantes son las semi-adosadas y las de segunda mano. Si el objetivo es construir la vivienda para venderla, la opción de viviendas adosadas son las que mayor estabilidad de precio presentan. Finalmente, si se tienen

viviendas de alquiler, se observa que los precios de los contratos bajan a final de año y verano, por lo que sería óptimo esperar a meses intermedios para renovarlos[3].

6. Recomendaciones de uso

Los resultados obtenidos indican que los modelos desarrollados son capaces de realizar predicciones adecuadas. Por tanto, surgen varias oportunidades de uso:

- En primer lugar, es posible identificar que áreas proyectan un crecimiento sostenido, además de otras áreas con precios más estables con las que mitigar el riesgo.
- También se recomienda identificar los tipos de propiedad con mayor retorno de inversión de forma más detallada, aunque ya hemos visto que las viviendas semi-adosadas y de segunda mano presentan un buen potencial de crecimiento. De igual forma, se ha identificado que las mejoras de eficiencia energética, especialmente al pasar de niveles inferiores (como D) a superiores (como C), incrementan el valor de la venta.
- Además, en el estudio de la evolución temporal del precio se ha encontrado que la correcta planificación de la renovación de contratos de alquiler, puede llegar a mejorar drásticamente los precios. De esta forma, es posible generar unos ingresos continuos y aprovechar las predicciones estacionales de los precios.

Para expandir las posibilidades de uso del modelo, se pueden tomar varias aproximaciones:

- Por un lado, la actualización de los datos censales permite reflejar mejor las condiciones socioeconómicas de cada área.
- Por otro lado, la integración de estos modelos en una interfaz de consulta permitiría obtener predicciones por distrito de manera dinámica.

[3] Además, como se mostró al inicio del documento en la Figura 5, el principal tipo de viviendas con contratos de alquiler son pisos y apartamentos.

Apéndice

Referencias

[1] W. O. Gibin, "Uk property price data 1995-2023-04," Disponible en https://www.kaggle.com/datasets/willianoliveiragibin/uk-property-price-data-1995-2023-04/data (19/10/2024).

[2] O. for National Statistics (UK), "Official census and labour market statistics," Disponible en https://www.nomisweb.co.uk/query/select/getdatasetbygeog.asp (19/10/2024).

[3] G. maps and data, "Mapa de reino unido," Disponible en https://gadm.org/download_country.html (19/10/2024).

[4] OpenCage, "Api de posición geográfica por nombre (versión de prueba)," Disponible en https://opencagedata.com/ (19/10/2024).

[5] Nominatim, "Api de posición geográfica por nombre (open source)," Disponible en https://nominatim.org/release-docs/develop/api/Overview/ (19/10/2024).

[6] G. de Reino Unido, "Descripción de las variables de los datos de kaggle," Disponible en https://www.gov.uk/guidance/about-the-price-paid-data (19/10/2024).

[7] "Google maps platform," Disponible en https://developers.google.com/maps/documentation/geocoding/overview?hl=es-419 (25/10/2024).