COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

GRADO EN INGENIERÍA EN TECNOLOGÍAS INDUSTRIALES

TRABAJO FIN DE GRADO

OPTIMIZACIÓN DE PRECIOS RESIDENCIALES PARA ALQUILER Y COMPRA

Autor: Marcos Monjardín de Aranda

Directora: Miren Telleria Ajuriaguerra

Madrid

Agosto de 2025

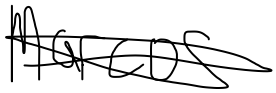
Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

OPTIMIZACIÓN DEL PRICING EN EL ALQUILER RESIDENCIAL

en la ETS de Ingeniería – ICAI de la Universidad Pontificia Comillas (Grado en Ingeniería en Tecnologías Industriales, Organización Industrial)
en el curso académico 2024/25 es de mi autoría, original e inédito y
no ha sido presentado con anterioridad a otros efectos.

El Proyecto no es plagio de otro, ni total ni parcialmente, y la información que ha sido tomada de otros documentos está debidamente referenciada.

Fdo.: Marcos Monjardín de Aranda **Fecha:** Agosto de 2025



Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Fdo.: Miren Telleria Ajuriaguerra **Fecha:** Agosto de 2025

COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

GRADO EN INGENIERÍA EN TECNOLOGÍAS INDUSTRIALES

TRABAJO FIN DE GRADO

OPTIMIZACIÓN DE PRECIOS RESIDENCIALES PARA ALQUILER Y COMPRA

Autor: Marcos Monjardín de Aranda

Directora: Miren Telleria Ajuriaguerra

Madrid

Agosto de 2025

OPTIMIZACIÓN DE PRECIOS RESIDENCIALES PARA PISOS DE ALQUILER Y COMPRA

Autor: Monjardín de Aranda, Marcos.

Director: Telleria Ajuriaguerra, Miren.

Entidad Colaboradora: ICAI – Universidad Pontificia Comillas

RESUMEN DEL PROYECTO

Este Trabajo de Fin de Grado analiza el mercado inmobiliario de la zona Nuevos Ministerios–Ríos Rosas (Madrid), con el objetivo de comprender qué factores determinan el precio de la vivienda en alquiler y en compraventa, y cómo pueden modelizarse de manera eficiente. Para ello, se han recopilado datos de más de 350 viviendas a través de extracción manual y de la API de un portal inmobiliario, considerando variables estructurales como superficie, número de habitaciones, número de baños, planta, exterioridad, presencia de ascensor y plaza de aparcamiento. El análisis se ha estructurado en tres fases principales, exploración de datos, modelización mediante regresión lineal y segmentación por perfiles de vivienda con clustering K-means.

Los resultados muestran que el tamaño del inmueble (superficie y número de habitaciones) constituye el factor determinante en la formación del precio, mientras que otras variables, como la planta o la orientación exterior, tienen un efecto más moderado. En particular, el modelo de regresión lineal permite explicar gran parte de la variabilidad de los precios confirmando la relevancia de los atributos básicos. Por su parte, el análisis de clustering ha permitido identificar grupos homogéneos de viviendas que reflejan tipologías diferenciadas, desde pisos pequeños económicos hasta viviendas de lujo de gran metraje. Cabe destacar que, mientras la regresión no detecta un impacto estadísticamente significativo del aparcamiento, el algoritmo K-means sí segmenta claramente las viviendas en función de disponer o no de plaza de garaje.

Además del análisis técnico, el trabajo aporta una reflexión metodológica sobre las estrategias de recogida de datos, en mercados reducidos como el analizado, más que recomendar exclusivamente el uso de una API, resulta fundamental plantear un estudio más longevo en el tiempo. A corto plazo, la recolección manual puede ser incluso más ágil, pero en el largo plazo la automatización mediante programación permite recopilar y actualizar un mayor volumen de datos de forma sistemática y eficiente.

Como líneas futuras, se plantea la posibilidad de simular anuncios ficticios con precios estimados por el modelo y subirlos a portales inmobiliarios para comprobar el tiempo necesario hasta recibir interés o reserva, lo que permitiría validar en un entorno real la capacidad predictiva del modelo. Asimismo, podrían incorporarse variables adicionales como antigüedad, reforma o calidad constructiva, ampliar el estudio a otras zonas de Madrid y aplicar modelos más complejos de machine learning (árboles de decisión, random forest, XGBoost) para aumentar la precisión predictiva.

En conclusión, este trabajo demuestra cómo la combinación de análisis exploratorio, regresión y clustering ofrece una visión complementaria y robusta del mercado inmobiliario, identificando los determinantes clave del precio y abriendo la puerta a aplicaciones prácticas en valoración, estrategias de marketing inmobiliario y políticas de vivienda.

Palabras clave: Vivienda, Precio, Regresión, Clustering, Mercado inmobiliario, Madrid.

OPTIMIZATION OF RESIDENTIAL PRICES FOR RENTAL AND SALE HOUSING

Author: Monjardín de Aranda, Marcos.

Supervisor: Telleria Ajuriaguerra, Miren.

Collaborating Entity: ICAI – Universidad Pontificia Comillas

ABSTRACT

This Final Degree Project analyses the real estate market in the Nuevos Ministerios–Ríos Rosas area (Madrid), with the aim of understanding which factors determine housing prices in both rental and sales markets, and how they can be efficiently modelled. For this purpose, data from more than 350 dwellings were collected through manual extraction and the API of a real estate portal, considering structural variables such as surface area, number of rooms, number of bathrooms, floor level, exteriority, the presence of an elevator, and parking availability. The analysis was structured in three main stages, exploratory data analysis, modelling through linear regression, and housing segmentation using K-means clustering.

The results show that the size of the dwelling (surface area and number of rooms) is the main factor in price formation, while other variables, such as floor level or exterior orientation, have a more moderate effect. In particular, the linear regression model explains a large share of the variability in prices, confirming the relevance of the basic attributes. Meanwhile, the clustering analysis identified homogeneous groups of dwellings reflecting differentiated typologies, ranging from small affordable flats to large luxury properties. It is noteworthy that while regression did not detect a statistically significant impact of parking, the K-means algorithm clearly segmented dwellings according to whether they included a parking space or not.

Beyond the technical analysis, the project also provides a methodological reflection on data collection strategies. In small markets such as the one analysed, rather than recommending the exclusive use of an API, it is essential to design a longer-term study that captures the dynamics of the market more comprehensively. In the short term, manual collection may even be more agile, but in the long run, automation through programming allows for the systematic and efficient gathering and updating of a greater volume of data.

As future work, the possibility is raised of simulating fictitious advertisements with prices estimated by the regression model and uploading them to real estate portals in order to observe

the time required to generate interest or a reservation, thereby validating the model's predictive capacity in a real environment. Furthermore, additional variables such as age of the building, renovation status, or construction quality could be incorporated, the study could be extended to other areas of Madrid, and more complex machine learning models (decision trees, random forest, XGBoost) could be applied to improve predictive accuracy.

In conclusion, this project demonstrates how the combination of exploratory data analysis, regression, and clustering provides a complementary and robust view of the real estate market, identifying the key determinants of price and opening the door to practical applications in valuation, real estate marketing strategies, and housing policies.

Keywords: Housing, Price, Regression, Clustering, Real estate market, Madrid.

Índice

Índice Ilustraciones

1. Introducción

1.1 Motivación

La motivación principal de este Trabajo de Fin de Grado ha sido aplicar herramientas de análisis de datos y modelos predictivos a un problema con impacto real en el entorno urbano, la determinación de precios en el mercado residencial, tanto en alquiler como en compraventa. En un contexto como el de Madrid, donde el precio de la vivienda ha mostrado un crecimiento sostenido y barrios como Ríos Rosas presentan una elevada rotación de inquilinos y dinamismo en las operaciones de compraventa, contar con estrategias de valoración basadas en datos se convierte en un elemento clave para mejorar la rentabilidad de propietarios, inversores y agentes, así como para favorecer la accesibilidad de inquilinos y compradores.

Este proyecto ha permitido poner en práctica los conocimientos adquiridos a lo largo del grado, en particular en el ámbito del análisis estadístico, la modelización de datos y la toma de decisiones fundamentadas en la evidencia. A diferencia de los métodos tradicionales basados en la intuición o en criterios subjetivos, el enfoque desarrollado ha integrado análisis exploratorio, modelos de regresión lineal y técnicas de segmentación mediante clustering K-means, ofreciendo una base objetiva que facilita la fijación de precios de forma más ajustada a las condiciones reales del mercado.

Además, el trabajo ha aportado una reflexión metodológica sobre las formas de recopilar información en mercados reducidos como el analizado. A corto plazo, la recogida manual de datos puede resultar más ágil, mientras que en horizontes temporales más largos la automatización mediante programación se plantea como la vía más eficiente para ampliar y actualizar la muestra. Este aspecto refuerza la importancia de diseñar estudios longitudinales que permitan capturar de manera más precisa la evolución del mercado en el tiempo.

En conjunto, el proyecto contribuye a mejorar la eficiencia y la transparencia del mercado inmobiliario promoviendo una valoración más racional y equitativa, con ello se pretende reducir desequilibrios entre oferta y demanda, acortar periodos de vacancia o de sobrevaloración y favorecer una asignación más equilibrada de recursos en zonas de alta presión inmobiliaria. La elección de Ríos Rosas como caso de estudio ha resultado adecuada al ofrecer un entorno representativo, con diversidad tipológica y suficiente actividad transaccional, que ha permitido llevar a cabo una investigación rigurosa y con proyección hacia futuros desarrollos.

1.2 Alineación con los ODS

Este proyecto se alinea con varios Objetivos de Desarrollo Sostenible de la Agenda 2030 de las Naciones Unidas, en tanto que busca mejorar la eficiencia, la transparencia y la equidad en el mercado del alquiler y compraventa residencial mediante el uso de datos y modelos de análisis:

ODS 8 – Trabajo Decente y Crecimiento Económico: Al promover prácticas de fijación de precios más racionales, el proyecto contribuye a la profesionalización del sector inmobiliario, especialmente en el ámbito de pequeños propietarios o gestores locales. La eficiencia en la gestión fomenta la inversión sostenible en vivienda y mejora la competitividad del mercado.

ODS 9 – Industria, Innovación e Infraestructura: El uso de tecnologías de análisis de datos, modelos predictivos y herramientas digitales en un sector tradicionalmente guiado por la intuición representa un paso hacia la innovación en el ámbito inmobiliario. Este trabajo pone en valor el papel de la digitalización para modernizar la toma de decisiones en el sector residencial.

ODS 10 – Reducción de las Desigualdades: Aunque de forma indirecta, una política de precios más objetiva y transparente puede contribuir a reducir desigualdades en el acceso a la vivienda, especialmente en zonas tensionadas, evitando situaciones de discriminación o precios desalineados con las características reales de los inmuebles.

ODS 11 – Ciudades y Comunidades Sostenibles: El objetivo principal del trabajo es contribuir a la sostenibilidad urbana mediante una fijación de precios más justa y basada en datos reales, ayudando a equilibrar la oferta y la demanda de vivienda en una zona de alta presión inmobiliaria como Ríos Rosas. Una mejor adecuación de los precios puede facilitar el acceso a la vivienda, reducir el tiempo de vacancia y evitar la especulación.

ODS 12 – Producción y Consumo Responsables: El proyecto fomenta un uso más eficiente de los recursos habitacionales disponibles, ya que permite ajustar mejor la oferta a la demanda real. Al reducir tiempos de vacancia y evitar precios inflados que distorsionan el mercado, se promueve un consumo de vivienda más racional y responsable, alineado con los principios de eficiencia y sostenibilidad.

2. Estado del Arte

2.1 Introducción al estudio del precio de la vivienda

El estudio del precio de la vivienda ha cobrado una creciente relevancia en el contexto económico y social, especialmente en países como España, donde el acceso a la vivienda constituye una de las principales dificultades para los hogares. Tanto la vivienda en propiedad como en régimen de alquiler representan no solo un bien de primera necesidad, sino también un activo financiero de gran peso en la economía doméstica y nacional. Comprender qué factores explican su precio y cómo pueden predecirse sus variaciones se ha convertido en una línea de investigación de gran interés para el ámbito académico, las instituciones públicas, las empresas y la sociedad en general.

El mercado inmobiliario cumple un doble papel, canaliza una parte importante del ahorro y la inversión privada, y al mismo tiempo influye en variables clave como el consumo de los hogares, la deuda, la estabilidad financiera y el bienestar social. Su evolución en España ha mostrado una elevada volatilidad en las últimas décadas, marcada por fases expansivas y contractivas vinculadas a cambios macroeconómicos, demográficos y normativos. La crisis financiera de 2008, la posterior recuperación y las alteraciones derivadas de la pandemia de COVID-19 son ejemplos recientes de cómo factores externos condicionan los precios de compra y alquiler, configurando un mercado complejo y sensible a choques de diversa naturaleza.

En los últimos años el aumento sostenido del precio del alquiler en ciudades como Madrid, Barcelona o Valencia ha reavivado el interés en analizar sus determinantes, especialmente en un contexto de creciente presión turística, procesos de gentrificación y una oferta limitada de vivienda pública. Estas dinámicas han generado debates sobre la necesidad de regulación, la creación de índices de referencia y el diseño de políticas activas de vivienda que garanticen un acceso más equitativo al mercado residencial.

Desde un punto de vista técnico, el análisis del precio de la vivienda puede abordarse a través de distintos enfoques. En primer lugar, los métodos tradicionales de valoración inmobiliaria, como la comparación, la capitalización de rentas o el coste de reposición, han servido históricamente como base de la práctica profesional. En segundo lugar, los modelos econométricos han permitido explicar de forma más precisa la influencia de los atributos estructurales, espaciales y socioeconómicos sobre los precios, incorporando además dimensiones temporales y geográficas. Finalmente, las nuevas metodologías de machine learning ofrecen herramientas con gran capacidad predictiva, capaces de procesar grandes

volúmenes de datos y capturar relaciones complejas que escapan a los modelos más convencionales.

En este marco, el presente trabajo se centra en revisar críticamente los principales enfoques teóricos y metodológicos empleados en la literatura para explicar y predecir el precio de la vivienda, poniendo especial atención en la interacción entre características físicas, localización y contexto socioeconómico, así como en la evolución reciente de las técnicas aplicadas al análisis inmobiliario.

2.2 Teoría de la valoración inmobiliaria

La valoración inmobiliaria es el proceso mediante el cual se estima el valor de un bien inmueble en un momento determinado, utilizando información técnica, económica y de mercado. Esta actividad es fundamental tanto para transacciones privadas como para la toma de decisiones de inversión, fiscalidad, seguros o financiación. A lo largo del tiempo se han consolidado diversos enfoques teóricos para valorar inmuebles cada uno con su propia lógica y ámbito de aplicación.

2.2.1 Fundamentos del valor inmobiliario

El precio de la vivienda, según la teoría económica clásica, surge del equilibrio entre la oferta y la demanda, pero presenta particularidades propias del mercado inmobiliario. La vivienda es a la vez un bien de consumo, que proporciona servicios habitacionales, y un activo de inversión, que genera rentas o plusvalías. El enfoque de activos de Poterba (1984) sostiene que el precio de compraventa se vincula al mercado de alquiler mediante el concepto de coste de uso, de forma que en equilibrio debe igualar el valor presente de los flujos futuros de renta esperados. Factores como los tipos de interés, las expectativas de revalorización y la fiscalidad inciden en la disposición a pagar, mientras que en la oferta pesan la heterogeneidad de cada inmueble, la escasa liquidez del mercado, la rigidez de corto plazo para aumentar el stock y la limitación del suelo disponible, especialmente en grandes ciudades.

A estos determinantes se añaden factores demográficos, económicos y cualitativos, como el crecimiento poblacional, el nivel de ingresos, la reputación del barrio o el entorno urbano. También influyen comportamientos especulativos que, en contextos de expectativas alcistas, pueden elevar los precios más allá de lo justificado por los fundamentales, como ocurrió en la burbuja inmobiliaria previa a 2008. Los mercados de compraventa y alquiler están

interrelacionados: un aumento de las rentas incrementa la rentabilidad de invertir en vivienda y puede trasladarse a mayores precios de venta, aunque en el corto plazo pueden darse desajustes por fricciones, regulación o choques específicos.

2.2.2 Métodos tradicionales de valoración

En la práctica profesional de la tasación inmobiliaria se han consolidado históricamente tres grandes enfoques, el método de comparación, el método de capitalización de rentas y el método del coste. Las Normas Internacionales de Valoración (IVSC, 2005) recogen estos tres métodos y establecen que todos parten del principio de sustitución, un comprador informado no pagará por un activo más de lo que costaría obtener otro de similares características por medios alternativos.

El **método comparativo** determina el valor de un inmueble a partir de los precios de mercado de propiedades similares vendidas o alquiladas recientemente. Para su aplicación se recopilan datos de transacciones recientes y se realizan ajustes para homogeneizar las diferencias entre el bien tasado y los comparables, considerando aspectos como superficie, estado de conservación, ubicación, antigüedad o equipamientos. Es el método más utilizado en la valoración de vivienda en España debido a su realismo y a la disponibilidad creciente de datos a través de registros y portales inmobiliarios. De hecho, la Orden ECO/805/2003 lo contempla como el método principal para determinar el valor de mercado en operaciones con garantía hipotecaria. No obstante, su precisión depende de la existencia de información suficiente y fiable. En mercados poco líquidos con escasa rotación de inmuebles o en el caso de propiedades muy singulares, puede ser difícil encontrar comparables adecuados, por lo que se complementa con otros enfoques.

El **método de capitalización de rentas** valora el inmueble en función de su capacidad para generar ingresos, generalmente mediante alquiler. Considera la propiedad como un activo de inversión cuyo valor corresponde al rendimiento económico que proporcionará a lo largo del tiempo. Puede aplicarse de forma directa, dividiendo el ingreso neto anual por una tasa de capitalización o mediante el descuento de flujos de caja futuros, empleando una tasa que refleje el coste de oportunidad del capital y el riesgo asociado. Este enfoque es especialmente útil para inmuebles en explotación, como viviendas en alquiler, edificios de apartamentos o locales comerciales, y conecta con la visión de la vivienda como activo financiero. La tasa de capitalización se obtiene del mercado o mediante modelos financieros, y su variación responde a cambios en los tipos de interés, en las primas de riesgo o en las expectativas de rentas. Aunque en mercados de inversión profesional es muy usado, su aplicación en la vivienda

habitual es más limitada, ya que muchas propiedades no generan renta de forma directa y su valor para el propietario incluye también la utilidad de uso.

El **método del coste** (o de reposición) estima el valor de un inmueble como la suma del valor del suelo más el coste actual de construir una edificación equivalente, descontando la depreciación acumulada por antigüedad, uso u obsolescencia. Este método parte de la idea de que un comprador racional no pagará por una vivienda más de lo que costaría edificarla de nuevo en su estado actual. Es especialmente relevante para inmuebles singulares o sin un mercado activo de comparables, así como para valoraciones con fines aseguradores. También proporciona un valor de referencia útil para contrastar con otros métodos, por ejemplo si los precios de mercado superan ampliamente el coste de reposición, puede indicar un componente especulativo, si caen por debajo, la construcción nueva puede desincentivarse, y así hasta que el equilibrio se recupere. Entre sus limitaciones se encuentra la dificultad de estimar con exactitud los costes de construcción y la depreciación, ya que requieren información actualizada y conocimiento técnico especializado.

Por último, están el método residual y el método del beneficio, son más útiles para inmuebles usados en otro ámbito que no sea vivienda. El método residual estima el valor del suelo en función del valor final de un proyecto inmobiliario, deduciendo del mismo los costes de construcción, honorarios, impuestos y margen del promotor. Es habitual en el análisis de promociones y suelo urbanizable, aunque requiere suposiciones fuertes sobre precios futuros. El método del beneficio se aplica principalmente a inmuebles vinculados a actividades económicas (hoteles, centros comerciales), en los que el valor está relacionado con la capacidad del inmueble para generar beneficios netos. Parte del análisis de la cuenta de resultados del negocio que ocupa el inmueble.

2.2.3 Limitaciones de los métodos tradicionales

Aunque consolidados, los métodos tradicionales presentan limitaciones relevantes. Requieren una cantidad considerable de datos fiables, suponen mercados estables y a menudo no capturan bien fenómenos como la segmentación espacial, las externalidades o las dinámicas temporales. Además, pueden ser sensibles a juicios subjetivos del tasador (por ejemplo, en ajustes comparativos) y muestran dificultades para adaptarse a contextos de alta variabilidad o cambio acelerado, como los que introduce el alquiler turístico o la presión gentrificadora.

Estas limitaciones han incentivado en los últimos años la incorporación de técnicas econométricas y algoritmos de aprendizaje automático (machine learning), que permiten

capturar relaciones no lineales, realizar análisis masivos y generar modelos predictivos con alta precisión.

2.3 Determinantes del precio de la vivienda

El precio de una vivienda no se define únicamente por su superficie o ubicación, sino que es el resultado de una combinación compleja de variables estructurales, espaciales y socioeconómicas que interactúan en el mercado inmobiliario. Comprender estos determinantes es esencial tanto para realizar valoraciones precisas como para identificar patrones o desigualdades en el acceso a la vivienda.

2.3.1 Variables estructurales

Las variables estructurales constituyen el punto de partida en cualquier análisis del valor de un inmueble, entre ellas destacan la superficie útil o construida, que en términos generales está positivamente relacionada con el precio absoluto, aunque no siempre implica un mayor valor por metro cuadrado. El número de habitaciones y baños se asocia directamente con la funcionalidad del inmueble y, en consecuencia, con la disposición a pagar de los potenciales demandantes. La antigüedad del edificio suele tener un efecto dual, mientras que las viviendas nuevas se valoran por la calidad de los materiales y las instalaciones, en determinados mercados los inmuebles antiguos rehabilitados pueden alcanzar precios igualmente altos. También influyen la altura y la orientación, ya que las plantas superiores con buena entrada de luz natural o vistas despejadas se valoran más, y la presencia de elementos adicionales como ascensor, terraza o garaje suele incrementar de forma significativa el precio. De hecho, estudios empíricos realizados en ciudades como Zaragoza u Oviedo han mostrado que variables como la existencia de ascensor o un mayor número de baños han sido algunos de los determinantes estructurales con mayor impacto en el precio de la vivienda en alquiler.

2.3.2 Variables espaciales

Las variables espaciales son igualmente determinantes puesto que el valor de la vivienda no depende solo de sus características internas, sino también del entorno urbano en el que se ubica. La literatura ha demostrado de forma consistente que las viviendas situadas en zonas céntricas, con buena accesibilidad al transporte público y proximidad a servicios esenciales, presentan precios más elevados. La distancia al centro urbano se ha consolidado como uno de los factores más influyentes, en línea con la teoría de la renta bidimensional urbana, que

explica la prima asociada a la localización central. Asimismo, la cercanía a estaciones de metro, líneas de cercanías o paradas de autobús aumenta el atractivo de una vivienda, al igual que la proximidad a colegios, centros sanitarios, zonas verdes o equipamientos culturales. Otros factores menos tangibles, como la reputación del barrio, la seguridad percibida o el grado de gentrificación, también juegan un papel importante en la formación de precios. En Oviedo, por ejemplo, se ha comprobado que el distrito en el que se ubica la vivienda influye de manera decisiva en el precio del alquiler, llegando en algunos casos a superar la incidencia de ciertas variables estructurales.

2.3.3 Factores socioeconómicos y de contexto

Los factores socioeconómicos y de contexto añaden otra capa de complejidad al análisis, la renta media de los hogares de la zona constituye un condicionante directo de los precios, ya que determina la capacidad de pago de los demandantes y condiciona la oferta disponible. La tasa de desempleo y el nivel educativo funcionan como indicadores indirectos de la estabilidad económica y la presión de demanda en un área determinada. A esto se suman aspectos demográficos, como la proporción de jóvenes, personas mayores, inmigrantes o estudiantes que influyen en el perfil de demanda y, por tanto, en el nivel de precios. En ciudades como Madrid, Barcelona o Valencia, el crecimiento del alquiler turístico, la gentrificación y la falta de vivienda pública han generado incrementos significativos en áreas concretas, elevando la presión sobre el mercado residencial. Un estudio reciente basado en redes bayesianas confirma que en Madrid los precios están más condicionados por factores relacionados con la accesibilidad y la intensidad de servicios, mientras que en Valencia predominan los elementos estructurales como el tamaño y la antigüedad de los inmuebles.

2.4 Modelos econométricos aplicados al mercado inmobiliario

El uso de modelos econométricos ha sido clave en el análisis del mercado inmobiliario, tanto para explicar cómo distintos factores inciden en el precio de la vivienda como para elaborar modelos predictivos o tomar decisiones de política pública. A diferencia de los métodos tradicionales de tasación, estos modelos permiten cuantificar el impacto de cada variable explicativa y establecer relaciones estadísticas robustas sobre grandes bases de datos.

2.4.1 Método de precios hedónicos

El método de precios hedónicos parte de una idea sencilla y es que el precio de una vivienda no depende de un único factor, sino de un conjunto de atributos que la hacen más o menos valiosa. Superficie, número de habitaciones, antigüedad, localización, calidad de los materiales, presencia de ascensor o incluso aspectos del entorno, como parques, colegios o zonas verdes, son elementos que el mercado incorpora al valorar cada inmueble.

A diferencia de otros métodos el enfoque hedónico no busca tasar un inmueble concreto caso por caso, sino analizar estadísticamente grandes muestras de datos para identificar cuánto aporta cada atributo al precio final. Esto se hace normalmente a través de modelos de regresión que permiten calcular precios implícitos de las características. Por ejemplo, se puede estimar cuánto aumenta el valor de un piso por tener una habitación más, por estar en una planta alta con ascensor o por encontrarse más cerca del centro urbano.

Este enfoque fue formalizado por Rosen (1974), que lo dotó de una base microeconómica sólida. Desde entonces se ha convertido en una de las herramientas más utilizadas por investigadores y organismos estadísticos, ya que permite aislar y medir con precisión los efectos de cada característica y elaborar índices de precios que corrigen cambios en la composición del mercado. Gracias a los modelos hedónicos se han podido cuantificar fenómenos como la prima que pagan las familias por vivir en barrios con mejores colegios, zonas más seguras o viviendas energéticamente eficientes.

Sin embargo, el método tiene algunas limitaciones, requiere bases de datos amplias y detalladas, y los resultados pueden variar según la forma matemática elegida para la estimación. También pueden aparecer problemas técnicos como la correlación entre variables (por ejemplo, superficie y número de habitaciones suelen estar relacionadas), o sesgos derivados de qué viviendas se incluyen en la muestra. Además, al tratarse de un enfoque estático, no siempre capta bien los cambios dinámicos del mercado o las expectativas de los compradores.

En España se han realizado múltiples aplicaciones. Caridad y Brañas (1996), por ejemplo, aplicaron este método en Córdoba y demostraron cómo la superficie, la edad de la vivienda o el barrio influían de forma clara en el precio. Aguiló (2002) hizo lo mismo en Baleares, encontrando patrones similares. Estos estudios confirman intuiciones conocidas, como que los pisos céntricos o con ascensor se pagan más, mientras que la antigüedad resta valor, pero las traducen en cifras concretas que ayudan a fundamentar las valoraciones tradicionales con evidencia empírica.

En definitiva, el método de precios hedónicos aporta una visión complementaria a los enfoques clásicos, permite entender el mercado no solo como un todo, sino también como la suma de pequeños atributos que explican por qué una vivienda vale más que otra.

2.4.2 Modelos con datos de panel

Los modelos con datos de panel son una herramienta muy útil cuando se quiere analizar cómo evolucionan los precios de la vivienda a lo largo del tiempo en diferentes lugares, como pueden ser municipios o barrios. Su principal ventaja es que permiten combinar información temporal y espacial, es decir, observar no solo cuánto varían los precios en distintos territorios, sino también cómo cambian con los años.

Una de sus mayores fortalezas es que ayudan a controlar factores que no se ven directamente en los datos pero que influyen en los precios, por ejemplo, la calidad de vida general de un barrio, su atractivo turístico o la reputación de una zona suelen ser elementos difíciles de medir, pero al mantenerse relativamente constantes en el tiempo pueden controlarse en el modelo. Esto se consigue mediante dos aproximaciones, los efectos fijos, que suponen que cada unidad (ciudad, barrio) tiene características propias invariables que afectan a los precios, y los efectos aleatorios, que asumen que esas diferencias se reparten de forma más general entre las unidades.

Gracias a esta flexibilidad los paneles permiten estudiar el impacto de variables dinámicas, como la apertura de una nueva línea de metro, una reforma normativa en alquileres o la evolución de la renta media de los hogares, aislando esos efectos de los rasgos permanentes de cada lugar. Así se obtiene una visión más completa de qué factores explican realmente la evolución del precio de la vivienda.

Un buen ejemplo es el caso de Cataluña, donde se analizaron datos de municipios de más de 5.000 habitantes entre 2013 y 2017. El estudio mostró que los precios por metro cuadrado aumentaron con más fuerza en municipios del área metropolitana de Barcelona, caracterizados por buena conectividad y alta presión de demanda, frente a zonas del interior con menor dinamismo económico. Este tipo de análisis confirma cómo los efectos locales y estructurales condicionan la trayectoria de los precios y cómo la vivienda en territorios bien comunicados y con fuerte atracción de población tiende a valorarse más rápidamente.

En definitiva, los modelos de datos de panel aportan una perspectiva valiosa porque permiten observar el mercado de la vivienda en movimiento, no solo explican cuánto valen los inmuebles en un momento concreto, sino también cómo y por qué esos valores cambian de un lugar a otro a lo largo del tiempo.

2.4.3 Modelos espaciales

El mercado de la vivienda no solo está condicionado por las características propias de cada inmueble o por factores económicos generales, sino también por su localización en relación con otras viviendas cercanas. En la práctica, los precios en un barrio o ciudad suelen estar correlacionados con los de zonas próximas, un aumento de precios en un distrito céntrico puede extenderse hacia barrios colindantes, generando lo que se conoce como efecto contagio.

Para capturar estas dinámicas, se emplean los modelos econométricos espaciales, que incorporan explícitamente la relación entre precios en distintas localizaciones. Existen principalmente dos enfoques:

- Modelos SAR (Spatial Autoregressive), que incluyen en la ecuación de precios la influencia de los valores observados en zonas vecinas.
- Modelos SEM (Spatial Error Models), que corrigen la dependencia espacial en el término de error, es decir, cuando los factores no observados en una zona están relacionados con los de áreas cercanas.

Gracias a estas técnicas, se pueden identificar patrones de concentración geográfica, como islas de precios altos o bajos, y entender cómo se transmiten las tensiones del mercado. Por ejemplo, en grandes ciudades, un incremento en la demanda en el centro suele generar aumentos de precios en barrios periféricos a medida que los compradores buscan alternativas más asequibles.

El uso de estos modelos es especialmente útil para la planificación urbana y las políticas de vivienda, ya que permite detectar zonas con riesgo de sobrecalentamiento, gentrificación o infravaloración. Además, cuando se combinan con datos de panel, ofrecen una visión aún más completa al analizar simultáneamente la dimensión espacial y temporal de los precios.

En resumen, los modelos espaciales ayudan a reconocer que el valor de una vivienda no se explica solo por sí misma, sino también por lo que ocurre alrededor. Incorporar la interacción geográfica en el análisis permite entender mejor la dinámica real del mercado inmobiliario, donde la proximidad y la conexión entre zonas juegan un papel clave en la evolución de los precios.

2.5 Nuevas metodologías basadas en machine learning

El desarrollo de técnicas de machine learning (ML) y el acceso a grandes volúmenes de datos (Big Data) han introducido nuevas posibilidades en la valoración y predicción del precio de

la vivienda. A diferencia de los modelos econométricos tradicionales que requieren especificar a priori la forma de las relaciones entre variables, los algoritmos de ML se centran en la capacidad predictiva y permiten identificar patrones no lineales e interacciones complejas en bases de datos extensas. Esto resulta especialmente relevante en el mercado inmobiliario, caracterizado por su alta heterogeneidad, abundancia de atributos espaciales y categóricos, y creciente disponibilidad de información procedente de registros, plataformas digitales o técnicas de web scraping.

2.5.1 Principales algoritmos empleados

Entre los algoritmos más utilizados en la literatura inmobiliaria destacan:

- Regresiones regularizadas (Ridge, Lasso), útiles como línea base para reducir sobreajuste y seleccionar variables relevantes.
- Random Forest, basado en ensambles de árboles de decisión, robusto frente a ruido y capaz de capturar relaciones no lineales.
- XGBoost, un modelo de *boosting* secuencial que ha mostrado un rendimiento superior al de métodos econométricos y estadísticos convencionales en múltiples estudios.
- Redes neuronales artificiales (ANN, DNN), con capacidad para aprender interacciones complejas en grandes bases de datos, si bien con menor interpretabilidad y mayor complejidad de ajuste.
- Otros enfoques como k-NN o SVM, aplicados en la predicción de índices nacionales de precios y con resultados comparables a técnicas más consolidadas.

2.5.2 Evaluación y límites

La comparación del desempeño entre modelos de ML y enfoques econométricos tradicionales se realiza habitualmente mediante métricas como el RMSE, el MAE o el R^2. En la mayoría de las aplicaciones los algoritmos de ML reducen de forma significativa los errores de predicción fuera de muestra, especialmente en mercados urbanos densos o cuando los precios responden a interacciones no lineales entre variables. Investigaciones recientes en Madrid y Terrassa han mostrado que Random Forest y XGBoost superan a la regresión lineal en precisión, mientras que estudios internacionales han confirmado la capacidad de XGBoost para alcanzar menores errores que incluso algunos métodos espaciales avanzados.

No obstante, el uso de ML plantea ciertos desafíos. El más destacado es la interpretabilidad, muchos algoritmos funcionan como “cajas negras”, dificultando justificar las valoraciones. Para mitigar esta limitación se emplean técnicas como SHAP o LIME, que permiten identificar la importancia de cada variable en las predicciones. Otro reto es la estabilidad de los modelos ya que pueden perder fiabilidad en escenarios de mercado distintos a los de entrenamiento, lo que obliga a su actualización constante.

2.5.3 Aplicaciones prácticas

Una de las aplicaciones más extendidas de ML en este ámbito son los Automated Valuation Models (AVMs), empleados por bancos, tasadoras y plataformas proptech. Estos sistemas integran grandes bases de datos de transacciones y características de viviendas, utilizando algoritmos como Random Forest, gradient boosting o redes neuronales para ofrecer estimaciones instantáneas del valor de los inmuebles. Aunque han logrado reducir considerablemente los errores medios frente a métodos tradicionales, la supervisión de expertos sigue siendo necesaria para casos atípicos o inmuebles con características singulares difíciles de cuantificar.

2.5.4 Diferencia entre modelos econométricos y modelos de machine learning

El estudio del precio de la vivienda se ha basado durante mucho tiempo en los modelos econométricos, cuyo objetivo principal es explicar de manera clara cómo influyen diferentes factores en el valor de un inmueble. Estos modelos, como los hedónicos, de panel o espaciales, parten de una estructura teórica definida y ofrecen resultados fáciles de interpretar: permiten saber, por ejemplo, cuánto aumenta el precio medio de una vivienda al añadir un baño o cómo influye la cercanía al centro urbano en su valoración. Su principal virtud es, por tanto, la transparencia y la capacidad de dar sentido a las relaciones entre las variables, aunque esta claridad suele implicar simplificaciones que no siempre capturan la complejidad del mercado.

En cambio, los modelos de machine learning se centran menos en explicar y más en predecir. No necesitan establecer de antemano cómo se relacionan las variables entre sí y son capaces de detectar patrones complejos en grandes volúmenes de información. Esta flexibilidad les permite incorporar datos muy variados, como los atributos de la vivienda, la localización, la evolución del mercado o incluso descripciones textuales e imágenes de los anuncios. La contrapartida es que sus resultados son menos fáciles de interpretar: aunque a menudo

predicen con mayor precisión que los modelos tradicionales, es más complicado entender con exactitud por qué el algoritmo ha asignado un determinado valor a una propiedad.

En definitiva, los modelos econométricos y los de machine learning no deben verse como enfoques opuestos, sino como perspectivas complementarias. Los primeros aportan claridad y comprensión, lo que resulta especialmente útil para el análisis académico o la elaboración de políticas públicas. Los segundos destacan por su capacidad predictiva y su adaptación a bases de datos masivas, lo que los convierte en herramientas cada vez más relevantes en el mercado inmobiliario digital. La combinación de ambos enfoques permite aprovechar lo mejor de cada uno, la explicación y la precisión, el rigor teórico y la potencia predictiva.

2.6 Conclusión: principales hallazgos y posicionamiento del trabajo

El análisis realizado muestra que el precio de la vivienda depende de una combinación de factores estructurales, espaciales y socioeconómicos. Los métodos tradicionales de valoración siguen siendo una referencia útil por su aplicabilidad práctica, aunque resultan limitados para explicar dinámicas más amplias del mercado. Los modelos econométricos han permitido identificar con rigor los principales determinantes del precio, como la superficie, el número de baños, la existencia de ascensor o la cercanía al centro urbano, así como factores de contexto como la renta de los hogares o la presión de demanda en determinadas zonas. Sin embargo, presentan dificultades para capturar relaciones más complejas o cambios en el tiempo y el espacio.

En este sentido, las técnicas de machine learning aportan un avance importante, ya que ofrecen una mayor capacidad predictiva y permiten aprovechar bases de datos amplias y variadas. Aun así, su menor transparencia y la dependencia de la calidad de los datos aconsejan utilizarlas como un complemento a los enfoques tradicionales más que como un sustituto.

A partir de estas conclusiones, este Trabajo de Fin de Grado se sitúa en la línea de investigación que combina metodologías, uniendo la capacidad explicativa de los modelos econométricos con la flexibilidad predictiva de los algoritmos de machine learning. La aplicación a un contexto local como Madrid, y en particular a la zona de Ríos Rosas–Nuevos Ministerios, permitirá comparar los resultados con los de otros mercados urbanos previamente estudiados y aportar una visión empírica propia sobre un área de especial relevancia en la capital.

Con ello, se busca no solo estimar con mayor precisión los factores que influyen en el precio de la vivienda, sino también contribuir al debate actual sobre el acceso a la vivienda en España,

ofreciendo evidencia útil para la toma de decisiones por parte de instituciones, empresas y ciudadanos.

3. Plan de Trabajo y Metodología

El trabajo se ha planteado como un estudio de caso sobre un micro mercado inmobiliario definido mediante un polígono trazado manualmente sobre el área comprendida entre Nuevos Ministerios y Ríos Rosas, recogiendo así las viviendas situadas dentro de los límites marcados por las principales calles del entorno. Esta acotación espacial persigue dos objetivos: por un lado, reducir la heterogeneidad en características urbanas y de demanda que aparece cuando se mezclan barrios con dinámicas distintas; y por otro, ganar precisión en la lectura de patrones locales de precio en venta y alquiler. De este modo, la zona opera como un laboratorio controlado en el que analizar ambos mercados bajo criterios consistentes.

En una primera fase, el mercado se delimitó mediante el polígono inicial. Sin embargo, al incorporar la API de Idealista fue necesario adaptar la estrategia, ya que esta herramienta exige definir un punto central y un radio de búsqueda. Por ello se tomó como referencia la dirección Alonso Cano 77 (Madrid) como centro y se fijó un radio de 500 metros, lo que permitió capturar un conjunto de datos que, aunque no reproduce de manera exacta el polígono inicial, se ajusta de forma representativa a la zona elegida y mantiene la coherencia con el planteamiento original del proyecto. La zona inicialmente elegida es la siguiente:

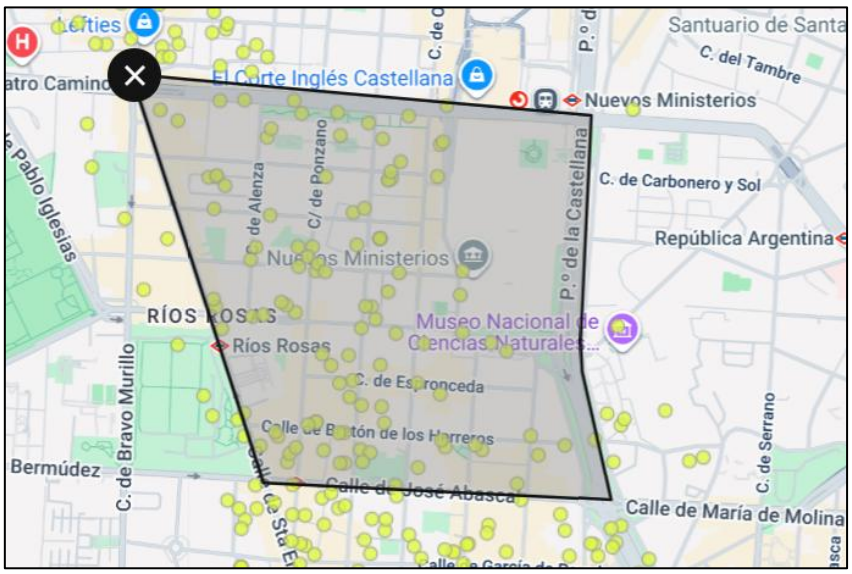


Ilustración 1: Área de estudio

3.1 Organización del trabajo

Fase 1. Recopilación manual y primer análisis

El proyecto comenzó con la extracción manual de información de pisos en venta dentro de la zona definida. Los anuncios seleccionados se registraron en hojas de Excel, conformando un conjunto inicial de 100 viviendas con variables fundamentales (precio, superficie, número de habitaciones, planta, localización, entre otras). Esta primera fase permitió familiarizarse con la estructura de la oferta y diseñar un esquema de análisis. Sobre ese esquema se implementó un primer código de análisis en Python que incluyó exploraciones descriptivas y una regresión lineal preliminar para estudiar relaciones básicas. Además, al tratarse de un proceso manual, fue posible seleccionar con precisión qué variables incorporar y razonar caso por caso aquellos anuncios en los que la información no estaba presentada de manera homogénea, completando o descartando registros según su coherencia. Gracias a ello, en este primer conjunto de datos se contó con un mayor número de variables inicialmente recopiladas y se presentaron menos dificultades en la posterior fase de limpieza, en comparación con los datos extraídos de manera automática.

Fase 2. Ampliación del alcance con la API

La concesión de acceso a la API de Idealista supuso un cambio significativo en el plan, al permitir la extracción automática de registros directamente desde la plataforma para venta y alquiler en el mismo perímetro. Se fijó como punto central Alonso Cano 77 y se mantuvo el radio de 500 metros para garantizar comparabilidad con la primera fase. La consulta devolvió 159 viviendas en alquiler y 108 en venta disponibles dentro del área. Se desarrolló un script específico para la extracción de datos, que posteriormente fueron volcados a hojas de Excel a efectos de control y trazabilidad, y a continuación integrados en el entorno de análisis, reutilizando la estructura de código previamente creada para el conjunto manual.

Fase 3. Integración y análisis paralelo

Con ambos orígenes (manual y API) el trabajo pasó a una fase de integración bajo un mismo esquema de variables para asegurar que el análisis se realizara sobre criterios homogéneos. La limpieza y estandarización detallada de los datos, tratamiento de duplicados, coherencia de rangos, normalización de categorías, se aborda en el apartado Extracción y preprocesamiento de datos. Aquí basta señalar que se aplicó un criterio de consistencia: cualquier transformación adoptada para venta se replicó para alquiler, y viceversa, con el fin de evitar sesgos de

procedimiento. A partir de esta base común se llevó a cabo el análisis de cada mercado por separado, aplicando idénticos supuestos en las relaciones exploradas.

Fase 4. Síntesis y preparación de resultados

La última fase del plan de trabajo se centró en reunir y ordenar los principales hallazgos, destacando los niveles de precio, la composición de la oferta y los patrones que aparecieron dentro del área analizada. Con ello quedó preparado el terreno para el capítulo de modelado y validación, en el que se comprobará con mayor detalle la solidez de dichas relaciones.

3.2 Criterios de selección

Se ha hecho una delimitación espacial, a través de la elección de un radio de 500 metros alrededor de Alonso Cano 77 persigue capturar un microentorno urbano homogéneo (trama, servicios, accesibilidad y tipologías constructivas relativamente similares) que reduzca la variabilidad ajena a las características intrínsecas de las viviendas. Esta aproximación de carácter nicho facilita que los cambios en precio reflejen con mayor probabilidad diferencias propias del activo (tamaño, estado, distribución) y no tanto efectos de barrio ni de localización en sentido amplio.

Como unidad de análisis está el anuncio de vivienda, por otro lado, la extracción manual y la automática se realizaron en la misma zona y en una ventana temporal próxima, lo que permite reducir el riesgo de sesgos derivados de la estacionalidad o de la rotación rápida de anuncios. En el capítulo de Extracción y preprocesamiento de datos se documentan los controles sobre fechas de captura y coherencia temporal.

Se han usado criterios de inclusión y exclusión, para empezar, se priorizaron viviendas residenciales con información suficiente en variables clave, cómo precio y superficie como mínimo. Aquellos registros con carencias esenciales, duplicados evidentes o incoherencias manifiestas se trataron siguiendo las reglas establecidas en el apartado técnico de datos. La premisa fundamental fue mantener un conjunto representativo del micro mercado sin sacrificar calidad ni comparabilidad.

3.3 Enfoque metodológico

El análisis se ha estructurado en torno a dos bloques diferenciados, el mercado de venta y el mercado de alquiler, cada uno de ellos se estudia de manera independiente, aplicando los mismos indicadores y técnicas para garantizar coherencia en los resultados. En una primera etapa se realizó un análisis descriptivo mediante medidas estadísticas básicas (medias, medianas, dispersión) y estudio de distribuciones de variables clave como precio, superficie, número de habitaciones o planta.

Posteriormente se exploraron relaciones bivariantes entre estas variables y el precio, con especial atención a la superficie y al número de dormitorios por su influencia en la formación del valor, a través de regresiones lineales sencillas planteadas como modelo base. De forma adicional, cuando fue pertinente se aplicaron segmentaciones por tramos de superficie o por características disponibles, siempre dentro de cada mercado por separado, con el objetivo de identificar patrones diferenciales en los precios.

Herramientas y trazabilidad, se combinaron Excel, para el registro y control inicial, verificación manual y exportaciones, y Python, para la extracción vía API, integración, análisis y generación de resultados. Para asegurar la reproducibilidad se mantuvo una separación clara entre datos brutos, datos preparados para análisis y salidas intermedias, junto con un registro de parámetros críticos, cómo punto central y radio, filtros aplicados y fecha de consulta.

3.4 Decisiones y alternativas

En el conjunto de datos se aprecia una diferencia en el número de registros entre viviendas en alquiler (158) y en venta (108). Esta situación es frecuente en micro mercados de reducidas dimensiones, donde la oferta disponible suele ser desigual entre ambos segmentos. Para no introducir sesgos, no se forzó un equilibrio artificial entre los grupos. En su lugar, se emplearon estadísticos robustos y, en los casos en que se presentan comparaciones de medias, se controló que la distinta distribución de tamaños no distorsionara la interpretación. Como alternativa metodológica se plantea la opción de trabajar con cuantiles de superficie, lo que permite analizar los resultados por tramos homogéneos de vivienda y facilita la detección de patrones representativos.

3.5 Calidad, sesgos y consideraciones éticas

El estudio se fundamenta en información pública de anuncios inmobiliarios, sin recoger ni tratar datos de carácter personal. En todo momento se han respetado el marco legal y las condiciones de uso de la plataforma, asegurando que las salidas del análisis mantengan la anonimización completa de las viviendas.

Desde la perspectiva metodológica, se han identificado algunos riesgos relevantes:
- Rotación y estacionalidad de los anuncios, el contenido de los portales inmobiliarios cambia con rapidez. Para mitigar este riesgo, la captura de datos se realizó en una ventana temporal acotada.

- Cobertura distinta según la fuente, la combinación de registros manuales y extraídos mediante API puede introducir discrepancias. Se aplicó un criterio de homogeneización de variables e inclusión coherente en ambos conjuntos.

- Valores atípicos o incoherentes, algunos casos extremos pueden distorsionar los resultados en un mercado reducido. Su tratamiento se aborda en Extracción y preprocesamiento de datos y, cuando procede, se muestran resultados con y sin dichos registros.

En todas las fases del trabajo se ha procurado mantener un enfoque transparente, documentando las decisiones adoptadas y reconociendo las limitaciones del método. El objetivo es asegurar que los resultados sean representativos, fiables y reproducibles.

3.6 Enlace con modelado y validación

Aunque el modelado y la validación se desarrollan en un capítulo posterior, conviene dejar definido el plan metodológico. Se emplearán regresiones lineales con el precio y el precio por metro cuadrado como variables dependientes, y un conjunto reducido de variables explicativas de alta disponibilidad y bajo riesgo de multicolinealidad. Estos modelos tendrán carácter exploratorio y servirán como base para evaluar la relación entre los factores clave del micro mercado.

Se prevé estimar modelos análogos tanto para el mercado de venta como para el de alquiler, con el fin de comprobar la significación de los coeficientes y la coherencia de los resultados bajo un mismo marco espacial. En cuanto a la validación, se aplicará un enfoque interno que incluirá el diagnóstico de residuos, la estabilidad de resultados por submuestras y, cuando el tamaño lo permita, una validación simple dividiendo los datos en un 85% de entrenamiento y un 15% de prueba (método holdout), lo que permitirá verificar el modelo con observaciones

no empleadas en la estimación. Asimismo, se contempla la realización de pruebas de robustez ante cambios de especificación, como la exclusión de valores atípicos evidentes o la inclusión de variables categóricas adicionales.

De este modo, el capítulo de modelado no parte de cero, sino que se enlaza de manera natural con el plan de trabajo y la metodología, reforzando la coherencia global del proyecto.

4. Extracción y Preprocesado de Datos

4.1 Fuentes de datos

La fuente de datos principal utilizada en este trabajo ha sido el portal inmobiliario Idealista, se trata de la plataforma más reconocida y consolidada en España en el ámbito de compraventa y alquiler de inmuebles, lo que asegura que los datos extraídos sean representativos y actualizados de la realidad del mercado. Idealista concentra una gran parte de la oferta inmobiliaria, cuenta con presencia nacional y ofrece un volumen de anuncios difícil de encontrar en otros portales competidores.

La elección de Idealista se fundamenta en su fiabilidad y en el hecho de que los anuncios incorporan información detallada y estructurada sobre cada vivienda, lo que resulta esencial para el objetivo de este trabajo. Otras fuentes potenciales, como Fotocasa o Habitaclia, fueron consideradas, pero presentan menor cuota de mercado y menor homogeneidad en el detalle de la información. Asimismo, fuentes oficiales como el Instituto Nacional de Estadística (INE) o el Catastro ofrecen datos más agregados, orientados a estudios macroeconómicos, por lo que no resultaban idóneas para construir un modelo de regresión lineal a nivel de inmueble.

No obstante, Idealista presenta limitaciones relevantes que conviene señalar. En primer lugar, aunque dispone de una API oficial, su acceso no es público ni inmediato. Requiere un proceso de registro y solicitud de credenciales, lo que supone una barrera para estudiantes o investigadores que buscan un acceso ágil a los datos. En algunos casos se han concedido permisos gratuitos para proyectos académicos, aunque esto no está garantizado y depende de la valoración de la solicitud. En cambio, las versiones profesionales de la API, orientadas a empresas inmobiliarias o entidades financieras, ofrecen funcionalidades avanzadas como métricas de zona, comparables o valoraciones automáticas, pero su coste económico es considerable y las condiciones de uso son estrictas. Esto explica por qué en la práctica la obtención de datos mediante esta vía resulta difícil y poco viable para un trabajo académico.

En segundo lugar, los anuncios publicados en Idealista pueden reflejar determinados sesgos del mercado, la plataforma recoge únicamente aquellas viviendas que son dadas de alta por particulares o agencias que deciden utilizar este portal, dejando fuera una parte significativa de la oferta que se canaliza por otras vías. Además, aunque Idealista pone a disposición de los usuarios herramientas de estimación de precios basadas en datos históricos y comparables, el precio final del inmueble es fijado siempre por el anunciante. Esto implica que en muchos casos los precios responden más a las expectativas o intereses del propietario o de la agencia que a un valor de mercado objetivo. En consecuencia, aunque los datos de Idealista son representativos de una parte muy relevante de la oferta inmobiliaria, no pueden considerarse un reflejo exhaustivo y neutral del conjunto del mercado.

Cabe destacar que, a dos días de la entrega, Idealista concedió acceso a su API oficial en el marco de este proyecto. Esto permitió complementar el proceso de recopilación con un método automatizado y más eficiente, especialmente en el ámbito de los datos de alquiler. Gracias a esta herramienta, se pudo trabajar con información estructurada y homogénea directamente desde la fuente, lo que mejoró la calidad del dataset y reforzó la validez del análisis posterior.

4.2 Método de extracción de datos

La extracción de datos no fue un proceso sencillo y constituyó una de las fases más laboriosas de todo el trabajo, durante varias semanas se invirtió un tiempo considerable en explorar la posibilidad de acceder a los datos de manera automatizada. En primer lugar, se contactó con Idealista para solicitar acceso a su API, sin obtener respuesta en el contexto de un proyecto académico. En vista de que no se obtendrá respuesta de Idealista, se investigaron también herramientas de web scraping como Octoparse o Lobstr.io, que habían sido empleadas en otros trabajos de fin de grado similares, pero las versiones gratuitas resultaron demasiado limitadas y las versiones de pago suponían un coste elevado que no se ajustaba a los recursos disponibles. Asimismo, se consultó a la empresa UrbanData Analytics, que facilitó un archivo Excel con información de inmuebles en alquiler en Chamberí, sin embargo, ese dataset resultaba demasiado general para los objetivos de este estudio, carecía de algunas variables clave y estaba agregado a un nivel espacial más amplio del deseado. Aun así, se conservó para realizar comparaciones exploratorias más generales en etapas posteriores y contrastar si daba resultados consistentes con el dataset principal.

Ante estas dificultades, se optó finalmente por realizar la extracción de manera manual. Se recopilaron un total de 100 anuncios de la zona de Nuevos Ministerios–Ríos Rosas,

considerada suficientemente homogénea como para reducir la variabilidad geográfica en los precios. Cada anuncio fue revisado y transcrito cuidadosamente en un archivo Excel, garantizando así la consistencia y fiabilidad del dataset construido.

Como previamente mencionado, Idealista concedió acceso a su API oficial, lo que permitió complementar el trabajo con un proceso automatizado de extracción de datos, en este caso centrados en el mercado de alquiler, este acceso supuso un cambio significativo en la metodología, al pasar de una recopilación manual limitada a un sistema estandarizado, escalable y más fácilmente reproducible. Gracias a la API, los datos se obtuvieron en formato estructurado (JSON), lo que aseguró una mayor homogeneidad en los registros y redujo considerablemente la probabilidad de errores de transcripción.

Para llevar a cabo este proceso, se desarrolló un script en Python que establecía conexión con la API mediante las credenciales proporcionadas por Idealista (API key y secret), a través de este código se podían lanzar consultas parametrizadas que permitían filtrar la información por criterios como localización geográfica, rango de precios, número de habitaciones o superficie. Las respuestas de la API eran convertidas de forma automática a un formato tabular e integradas en un DataFrame de Pandas, lo que facilitó su análisis posterior y la preparación de los datos para el modelo de regresión.

El uso de este procedimiento no solo incrementó la eficiencia en la recopilación de datos, sino que también mejoró su consistencia interna, por ejemplo, la estructura de la API garantizaba la presencia de campos obligatorios sin valores nulos y ofrecía variables categóricas y dicotómicas en un formato sencillo de transformar a binario. Esto redujo notablemente la necesidad de realizar procesos de limpieza extensos, en comparación con el dataset generado manualmente. En definitiva, la integración de la API de Idealista reforzó la solidez metodológica del trabajo, al proporcionar un mecanismo de extracción fiable y reproducible.

Se seleccionaron catorce variables iniciales: precio, superficie, número de habitaciones, número de baños, antigüedad, estado de reforma, planta, ascensor, aparcamiento, terraza, amueblado, aire acondicionado, piscina y exterioridad. Ocho de ellas se codificaron como dicotómicas (reformado, ascensor, aparcamiento, terraza, amueblado, aire acondicionado, piscina y exterior), mientras que el resto eran continuas (precio, superficie y antigüedad) o categóricas (habitaciones, baños y planta).

Durante el preprocesamiento se tomaron decisiones de simplificación para mejorar la utilidad del modelo, la variable antigüedad se eliminó debido a la dificultad de homogeneizar su codificación y a la escasa variabilidad observada en la muestra. Del mismo modo, otras variables como aire acondicionado o ascensor mostraban una distribución muy sesgada hacia

la respuesta afirmativa, con una mayoría de viviendas que sí disponían de dichas características, lo que reducía notablemente su capacidad explicativa dentro del modelo.

4.3 Estructuración inicial de los datos

La organización inicial del dataset se llevó a cabo en Excel, donde se configuró una estructura clara y sistemática, cada fila correspondía a un inmueble y cada columna a una de las variables seleccionadas. Este enfoque matricial resultaba idóneo para una posterior exportación a Python, ya que se adapta perfectamente al formato requerido por los algoritmos de regresión lineal, en los que cada variable actúa como predictor potencial del precio de la vivienda.

En esta fase, el principal objetivo era garantizar que la información quedara registrada de manera homogénea y sin ambigüedades, sin embargo, pronto surgieron algunas dificultades en la forma en que los datos aparecían en los anuncios. Un ejemplo claro fue el caso del precio, ya que en algunos anuncios se indicaba un precio base al que podían añadirse extras opcionales, como una plaza de garaje. Para evitar inconsistencias y reflejar un valor más realista de la transacción, se decidió que, cuando aparecía la opción de adquirir un garaje por un coste adicional, este importe se sumaría al precio total del inmueble, y simultáneamente se marcaría la variable de aparcamiento como presente (1). De este modo, se reflejaba no solo la existencia de la plaza, sino también el impacto económico que tendría en el valor total de la operación.

Asimismo, la variable de estado de la vivienda o reforma presentó una notable heterogeneidad en su descripción. Algunos anuncios señalaban explícitamente que la vivienda era “obra nueva”, otros la definían como “en buen estado” o “recién reformada”, mientras que en muchos casos se utilizaban términos más vagos como “a actualizar” o simplemente no se especificaba nada. Para homogeneizar esta variable se creó una codificación binaria en la que se asignaba un valor de 1 siempre que la descripción incluyera términos que denotaran renovación o conservación adecuada, y un valor de 0 en el resto de los casos.

En el caso del dataset obtenido mediante la API de Idealista se aplicaron verificaciones adicionales para confirmar que no existieran valores nulos ni outliers, y se eliminaron automáticamente los duplicados detectados en el proceso de extracción. Estas comprobaciones se realizaron mediante el desarrollo de rutinas de programación en Python, que permitieron revisar de forma sistemática la consistencia de todas las variables y garantizar que los registros incorporados al análisis fueran completos y fiables.

Del mismo modo, las variables de carácter dicotómico ya se transformaron en binarias (0/1) dentro de este proceso de preprocesamiento automático, de esta manera, características como

ascensor se codificaron directamente como presencia o ausencia, simplificando la preparación del dataset para su integración en el modelo de regresión. Este trabajo realizado a través de código contribuyó a reducir el riesgo de errores manuales y a asegurar que el conjunto de datos final presentase un formato coherente y adecuado para el análisis estadístico posterior.

En definitiva, esta etapa de estructuración permitió convertir la información, inicialmente dispersa y heterogénea, en un conjunto de datos organizado, pero puso de manifiesto la necesidad de un proceso más profundo de normalización que se abordó en la fase de limpieza.

4.4 Limpieza de datos

La limpieza de datos constituyó una fase crítica para garantizar la fiabilidad del análisis y la robustez del modelo posterior, en primer lugar, se procedió a identificar y eliminar anuncios duplicados, ya que algunos inmuebles aparecían más de una vez en el portal, bien por estrategias de las agencias inmobiliarias o por actualizaciones del anuncio. La depuración de estos registros evitó que se otorgara un peso desproporcionado a determinadas viviendas dentro del análisis.

En segundo lugar, se abordó el problema de los valores faltantes, en aquellas variables consideradas fundamentales para la construcción del modelo, como precio, superficie o número de habitaciones, los registros con información incompleta fueron descartados, dado que su inclusión podría distorsionar los resultados. En cambio, para las variables de carácter dicotómico (como ascensor, terraza o aire acondicionado), se adoptó el criterio de asumir que la ausencia de información equivalía a un valor negativo (0). Esta decisión se justificó en que, en la mayoría de los portales inmobiliarios, las características más atractivas suelen destacarse explícitamente, mientras que la omisión suele implicar su inexistencia.

Otro aspecto relevante fue la normalización de categorías textuales. Tal como se explicó en el apartado anterior, la variable de reforma fue recodificada en una escala binaria para eliminar ambigüedades, de igual manera, se establecieron criterios uniformes para variables como aparcamiento, de modo que si aparecía la opción de adquirirlo por un coste adicional, se consideraba que el inmueble disponía efectivamente de esa característica y se incorporaba el importe al precio final, como explicado anteriormente.

En cuanto a la superficie, se resolvió la heterogeneidad entre superficie útil y construida mediante la adopción sistemática de la superficie construida, garantizando así la comparabilidad entre todos los registros, se debe a que se entendió que es la variable favorable para publicar y que los anunciantes que no tuviesen ambas publicadas, tendrían superficie construida publicada y no superficie útil. Finalmente, en relación con los outliers, si bien se

identificaron algunos inmuebles con precios significativamente más altos que la media, se optó por mantenerlos en la base de datos. La razón fue doble, por un lado, eliminarlos podría haber reducido artificialmente la variabilidad natural del mercado, y por otro, se consideró que dichos valores extremos aportaban información relevante sobre el rango real de precios en la zona analizada, la discusión más detallada sobre su impacto se difirió al análisis de resultados.

En conclusión, el proceso de limpieza no solo depuró el dataset de inconsistencias y errores, sino que también supuso la toma de decisiones metodológicas importantes que afectaron a la calidad y representatividad del conjunto de datos final. Este rigor en la preparación de la información fue determinante para poder aplicar, en fases posteriores, técnicas de regresión lineal con garantías de validez y fiabilidad

4.5 Preprocesamiento para el modelo

Una vez finalizada la fase de limpieza, el conjunto de datos fue exportado desde Excel a Python para su posterior análisis y modelización, en este entorno se utilizaron diversas librerías especializadas que facilitaron la manipulación de datos, la construcción del modelo y la evaluación de su rendimiento. Entre ellas destacan NumPy y Pandas, que resultaron fundamentales para el tratamiento de datos en estructuras matriciales y la gestión eficiente de tablas. Para la estimación econométrica se recurrió a Statsmodels, mientras que Scikit-learn fue empleado tanto en la división de la muestra en subconjuntos de entrenamiento y test como en el cálculo de métricas de error. Adicionalmente, se incorporaron Matplotlib y Seaborn con el fin de representar gráficamente los resultados y facilitar la interpretación visual de los patrones del dataset.

El código inicial de importación de librerías se muestra en el siguiente fragmento:

```
import numpy as np
import pandas as pd
import statsmodels.api as sm
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
from statsmodels.stats.outliers_influence import variance_inflation_factor
import matplotlib.pyplot as plt
import seaborn as sns
```

Ilustración 2: Código de importación de librerias

El uso de la API resultó especialmente útil en este punto, dado que los datos ya se encontraban estructurados de manera más homogénea y con menor presencia de inconsistencias, lo que facilitó la transición hacia la etapa de modelización en Python.

En esta fase no fue necesario aplicar procesos adicionales de codificación, ya que las variables dicotómicas habían sido previamente transformadas a valores binarios (0/1) durante la estructuración en Excel, lo que sí se llevó a cabo fue el escalado y normalización de variables continuas, como la superficie o el precio, con el fin de homogeneizar sus magnitudes y mejorar la estabilidad numérica del modelo. Asimismo, el dataset se dividió en dos subconjuntos, un conjunto de entrenamiento, que representaba el 85 % de las observaciones y que fue empleado para la estimación de los parámetros, y un conjunto de test con el 15 % restante, utilizado para validar la capacidad predictiva del modelo y detectar posibles problemas de sobreajuste.

Este conjunto de pasos metodológicos aseguró que el modelo de regresión lineal se construyera sobre una base sólida, minimizando errores derivados de la heterogeneidad en las variables y garantizando que los resultados obtenidos fueran fiables y reproducibles.

4.6 Reflexión y conclusión

A pesar de la rigurosidad del proceso, la calidad de los datos presenta ciertas limitaciones que conviene señalar, en primer lugar, el tamaño reducido de la muestra (100 inmuebles) condiciona la potencia estadística del modelo. Además, la extracción manual puede implicar errores humanos en la transcripción, aunque se aplicaron revisiones sistemáticas para minimizar este riesgo.

Por otra parte, es importante subrayar que el dataset no busca reflejar la totalidad del mercado inmobiliario madrileño, sino que se centra deliberadamente en una zona concreta, Nuevos Ministerios–Ríos Rosas, con el objetivo de analizar de forma específica y homogénea cómo se comportan los precios en ese entorno. La aparente limitación de circunscribirse a un área reducida se convierte, en realidad, en una fortaleza metodológica, ya que permite estudiar con mayor detalle un mercado local concreto sin que la variabilidad geográfica introduzca ruido en los resultados.

Aun así, incluso dentro de esta zona, los anuncios recopilados pueden estar sesgados hacia los inmuebles más atractivos o con mejor publicidad, ya que son precisamente los que tienden a publicarse en plataformas como Idealista. En consecuencia, aunque el dataset es representativo de la oferta visible en dicho portal, no necesariamente incluye todas las transacciones efectivas que se producen en el área.

De cara a investigaciones futuras, sería recomendable complementar este enfoque con técnicas de scraping más avanzadas, el acceso a APIs privadas o la integración de fuentes oficiales (Catastro, registros notariales, estadísticas municipales). Estas mejoras permitirían contrastar los resultados obtenidos, enriquecer el dataset y aumentar la solidez de las conclusiones sin perder el foco local que caracteriza a este estudio.

El proceso de extracción y preprocesamiento de datos ha requerido un esfuerzo significativo, dada la ausencia de herramientas automáticas de acceso a los datos de Idealista. Sin embargo, la estrategia adoptada ha permitido construir un dataset limpio, coherente y suficientemente detallado para alimentar un modelo de regresión lineal.

La calidad del preprocesamiento garantiza que los resultados obtenidos en fases posteriores del trabajo no se vean contaminados por errores de formato, duplicados o inconsistencias. Aunque las limitaciones de tamaño y representatividad de la muestra son innegables, este trabajo sienta las bases para análisis posteriores más amplios y rigurosos, demostrando la importancia de un tratamiento cuidadoso de los datos desde su extracción hasta su preparación final para el modelado.

5. Análisis Exploratorio de Datos

5.1 Descripción del conjunto de datos

El objetivo de este apartado es explorar de forma descriptiva el conjunto de datos recopilado sobre el precio de la vivienda en la zona de Nuevos Ministerios–Ríos Rosas (Madrid), distinguiendo entre el mercado de venta y el de alquiler Para ello, se han obtenido datos mediante la API de un portal inmobiliario, conformando dos subconjuntos, uno de viviendas en venta y otro de viviendas en alquiler, todos ubicados en la zona de estudio. Las variables disponibles en ambos casos incluyen el precio (en euros), la superficie (en metros cuadrados), el número de habitaciones, el número de baños, la planta en la que se encuentra la vivienda, y variables indicadoras como la disponibilidad de ascensor, de aparcamiento, así como la exterioridad. En total, el subconjunto de venta cuenta con alrededor de 108 observaciones, mientras que el de alquiler contiene en torno a 158 viviendas recopiladas. Cabe señalar que otras variables potencialmente relevantes, como la antigüedad del edificio o el estado de reforma, no se han incluido en el análisis por no estar disponibles a través de esta fuente de datos.

En este contexto, cada variable seleccionada cumple un papel específico en la explicación del valor de un inmueble, el precio es la variable dependiente y central del análisis, pues todas las

demás características se estudian en relación con él. La superficie constituye uno de los factores más determinantes en la valoración, ya que el tamaño suele asociarse directamente a un mayor valor de mercado, aunque no siempre de forma proporcional. Relacionado con ello, el número de habitaciones refleja la capacidad funcional de la vivienda, siendo un indicador clave del tipo de inmueble, el número de baños también aporta información relevante sobre el nivel de comodidad, dado que contar con varios baños suele estar asociado a viviendas más amplias o de gama más alta.

Por otro lado, la planta en la que se ubica la vivienda incide en la luminosidad y en las vistas, atributos que repercuten en su valor, los pisos altos, especialmente los áticos, tienden a ser más demandados, mientras que los bajos suelen tener precios menores. Variables adicionales como la plaza de aparcamiento constituyen atributos muy valorados en entornos urbanos, ya que aportan comodidad y seguridad y suelen incrementar el valor de la vivienda. La orientación exterior o interior influye en la luminosidad y en la ventilación, siendo habitual que los pisos exteriores alcancen precios más altos que los interiores. Finalmente, la presencia de ascensor es determinante en edificios de varias plantas, ya que su ausencia penaliza de forma significativa el precio, sobre todo en viviendas situadas en alturas elevadas.

En conjunto, estas variables han sido escogidas no solo por su disponibilidad en la fuente de datos, sino porque son reconocidas tanto en la literatura como en la práctica del mercado inmobiliario como determinantes del valor de la vivienda. Este análisis exploratorio permite familiarizarse con la estructura básica de los datos y detectar patrones iniciales. Además, servirá como punto de partida para los modelos posteriores de regresión y segmentación, que buscarán cuantificar con mayor precisión la influencia de cada una de estas características en la fijación de precios. A continuación, se muestran las matrices de correlación para venta y alquiler con todas las variables añadidas.

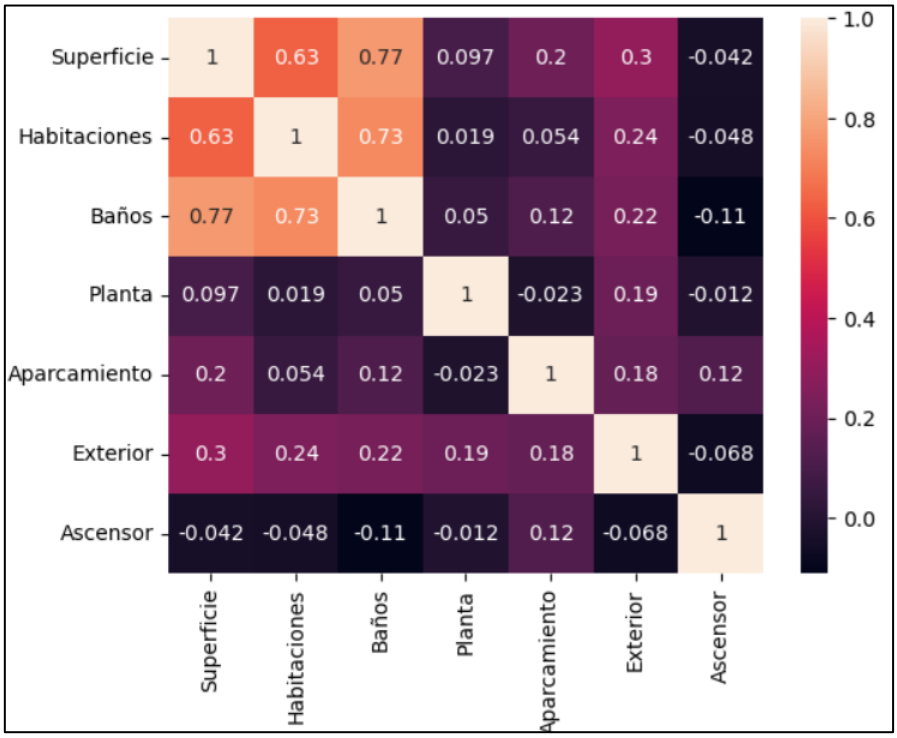


Ilustración 3: Matriz de correlación propiedades en alquiler

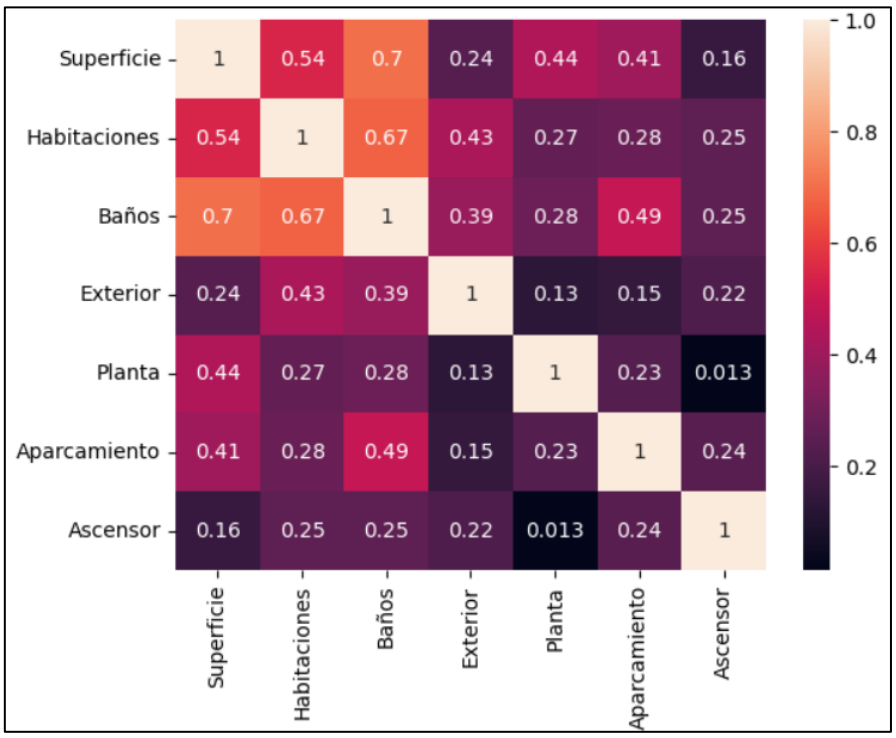


Ilustración 4: Matriz de correlación propiedades en venta (vía API)

5.2 Análisis univariante

5.2.1 Precio alquiler

En la ilustración encontrada en el apéndice se representa la distribución de los precios de alquiler mediante un diagrama de caja. El valor medio de las rentas en la muestra es de 2.600,06 € mensuales, la mediana se sitúa en torno a los 2.420 €, mientras que el primer y tercer cuartil se encuentran en 1.693,75 € y 3.200 €, respectivamente. Esto indica que el 50% central de los alquileres oscila en un rango relativamente amplio, entre aproximadamente 1.700 € y 3.200 €. Los valores mínimo y máximo muestran también una elevada dispersión. Se registran casos extremos por encima de los 5.800–6.500 €, que constituyen valores atípicos, aunque no dejan de reflejar la presencia de viviendas de alto nivel dentro de la zona de estudio. La dispersión hacia la parte superior es más marcada que hacia la inferior, lo que evidencia una asimetría positiva en la distribución de precios, existen algunas rentas muy altas que elevan la media, mientras que la mayor parte de la oferta se concentra en valores más contenidos.

En conjunto, este análisis inicial del precio de alquiler muestra un mercado heterogéneo, con gran parte de las viviendas situadas en torno a los 2.000–3.000 € mensuales, pero con una presencia destacada de inmuebles de lujo que alcanzan valores muy superiores.

5.2.2 Precio venta

En otra ilustración mostrada en el apéndice, se recoge la distribución de los precios de venta mediante un diagrama de caja. El valor medio en la muestra se sitúa en 1.474.580,37 €, lo que ya indica un mercado de precios elevados en la zona de Nuevos Ministerios–Ríos Rosas. La mediana aparece en torno a 1.135.000 €, con un rango intercuartílico comprendido entre 655.000 € (primer cuartil) y 1.637.500 € (tercer cuartil), es decir, la mitad central de las viviendas en venta se sitúa dentro de ese intervalo.

Los extremos de la distribución muestran una gran dispersión, el precio mínimo registrado en la muestra se aproxima a 199.000 €, correspondiente previsiblemente a viviendas de pequeño tamaño o de menor calidad relativa, en el otro extremo aparecen viviendas cuyo precio alcanza los 15.000.000 €, que constituyen valores atípicos muy superiores a la mayoría de la muestra. Entre los casos intermedios también se observan valores destacados por encima de los 4 millones de euros que reflejan la presencia de inmuebles singulares de lujo.

El gráfico evidencia que la distribución de precios de venta está fuertemente sesgada hacia la derecha, con la existencia de un número reducido de viviendas de muy alto precio que elevan de forma notable la media frente a la mediana. Este comportamiento es característico de mercados inmobiliarios de zonas céntricas y consolidadas, donde conviven viviendas de rango medio-alto con propiedades de lujo de precios extraordinariamente elevados.

5.2.3 Superficie alquiler

La caja y bigotes para esta sección muestra la distribución de la superficie de las viviendas en alquiler mediante un diagrama de caja, el valor medio de la muestra se sitúa en 93,01 m², mientras que la mediana se encuentra en torno a los 81 m². El rango intercuartílico oscila aproximadamente entre los 60 m² y los 117 m², lo que indica que la mitad de los pisos ofertados tiene una superficie comprendida en ese intervalo.

Los valores extremos reflejan la heterogeneidad del mercado, en la parte inferior el mínimo registrado es de 25 m², correspondiente previsiblemente a estudios o apartamentos muy reducidos. En la parte superior se observan casos atípicos con superficies que superan ampliamente la media, viviendas de 215 m², 250 m², 316 m² e incluso un caso excepcional de 540 m². Aunque constituyen outliers estadísticos, estos inmuebles responden a la existencia de pisos de gran tamaño en la zona generalmente vinculados a segmentos de lujo.

La distribución de la superficie presenta por tanto una clara asimetría positiva, con la mayor parte de la oferta concentrada en viviendas de tamaño medio, pero con una cola superior marcada por viviendas de gran metraje. Este patrón coincide con la realidad del mercado en barrios céntricos, donde conviven apartamentos compactos con pisos señoriales de dimensiones muy superiores.

5.2.4 Superficie venta

En esta sección, la caja y bigotes representa la distribución de la superficie de las viviendas en venta, el valor medio se sitúa en 154,56 m², mientras que la mediana alcanza aproximadamente los 116,5 m². El rango intercuartílico se extiende desde 75,75 m², hasta 182,75 m², lo que muestra que la mitad de la muestra se concentra en superficies de tamaño medio, comprendidas entre viviendas relativamente compactas y otras de mayor amplitud.

Los valores mínimo y máximo reflejan la elevada heterogeneidad del mercado, el inmueble más pequeño registrado presenta apenas 17 m², propio de estudios muy reducidos, mientras que en el extremo superior se observan viviendas con superficies excepcionales de 1.281 m² y hasta 1.500 m², que corresponden a propiedades singulares de gran lujo en la zona. También se detectan otros casos atípicos, como inmuebles de 426 m² o 459 m², que, si bien no alcanzan los extremos máximos, quedan claramente fuera del rango habitual.

El gráfico muestra, por tanto, una distribución con clara asimetría positiva, en la que predominan las viviendas de entre 75 y 180 m², pero donde existe un segmento reducido de propiedades de gran metraje que eleva notablemente la media respecto a la mediana. Este

patrón es característico de mercados céntricos con oferta diversificada, donde conviven apartamentos de tamaño medio con viviendas exclusivas de dimensiones muy superiores.

5.3 Análisis bivariante

En el análisis bivariante se examinan las relaciones entre el precio de la vivienda y algunas de las variables explicativas más importantes considerando por separado las viviendas en venta y en alquiler. En particular, se explora cómo varía el precio en función de la superficie, del número de habitaciones y de la planta. Para ilustrar estas relaciones se utilizan principalmente gráficos de dispersión y gráficos comparativos de medias, que permiten apreciar tendencias o patrones de correlación de forma visual.

5.3.1 Precio vs Superficie

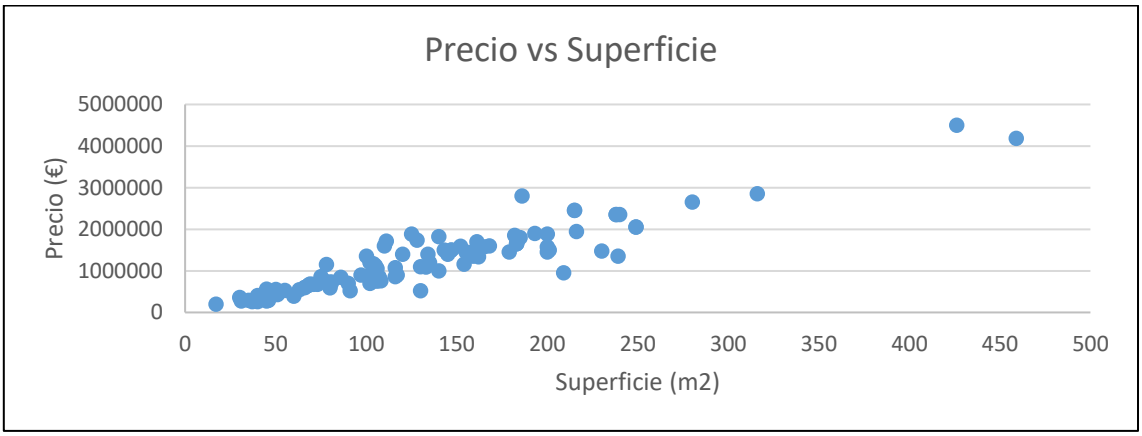


Ilustración 5: Relación entre precio de venta y superficie de la vivienda en Nuevos Ministerios–Ríos Rosas

En la Ilustración 5 se muestra un gráfico de dispersión del precio de venta frente a la superficie, se aprecia claramente una tendencia ascendente, conforme aumenta la superficie (m^2) de la vivienda, el precio de venta (en euros) también crece de manera notable. Los puntos dispersos siguen aproximadamente una línea de tendencia positiva, lo que indica que existe correlación entre ambas variables. Incluso aquellos inmuebles de superficie extraordinariamente grande (situados muy a la derecha del gráfico, tanto que no se ven ya que esta es una versión más corta para ver mejor la imagen, si se quiere ver el gráfico entero se puede encontrar en el apéndice) presentan precios muy altos acordes a su tamaño, en línea con la tendencia general. Por ello, aunque pudieran parecer valores atípicos por su magnitud, estos casos no se consideran anómalos en términos de la relación precio-superficie, sino que confirman el patrón esperado, una vivienda excepcionalmente grande conlleva un precio excepcionalmente alto. En resumen, en el mercado de venta la superficie emerge como uno de los factores

principales asociados al precio, las viviendas más amplias alcanzan precios significativamente mayores, lo cual es congruente con la intuición y con la realidad del mercado inmobiliario en la zona estudiada.

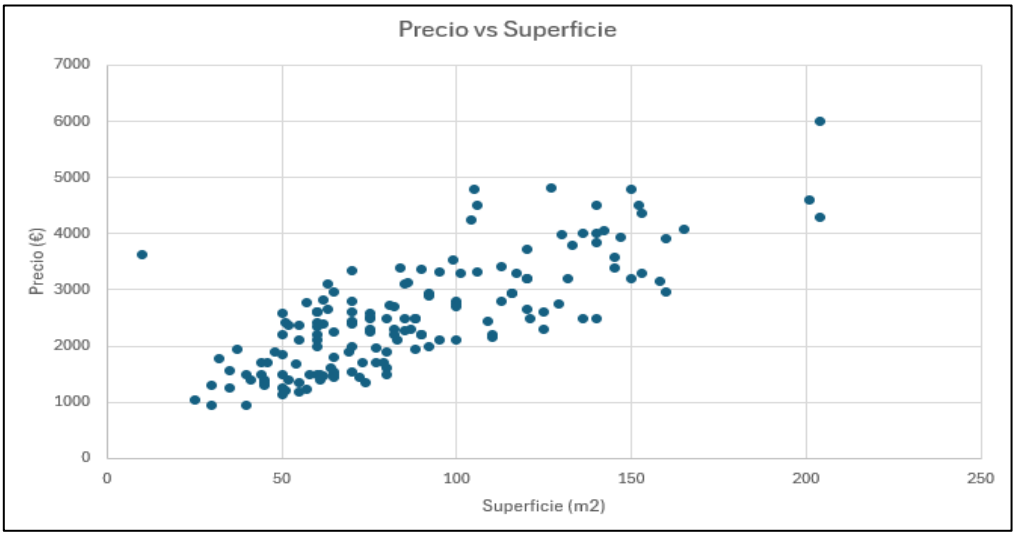


Ilustración 6: Relación entre precio de alquiler y superficie de la vivienda

De forma análoga, en el caso de las viviendas en alquiler se observa también una relación positiva entre el tamaño y el precio, la Ilustración 6 muestra que los pisos con mayor superficie suelen tener rentas mensuales superiores. Para facilitar la visualización, se han excluido del gráfico algunos casos de superficies extremadamente grandes, que en la muestra original superaban ampliamente al resto (cómo previamente mencionado), estos casos podrían considerarse outliers en términos de superficie, aunque sus rentas siguen la misma línea de tendencia creciente, y por eso se ha decidido mantenerlos para los siguientes análisis. Aun sin esos puntos de muy alta superficie la tendencia general persiste, los alquileres aumentan con los metros cuadrados. En definitiva, tanto en venta como en alquiler, la superficie de la vivienda es un factor claramente ligado al precio siendo las viviendas más espaciosas notablemente más caras en cada ámbito.

5.3.2 Precio vs Número de habitaciones

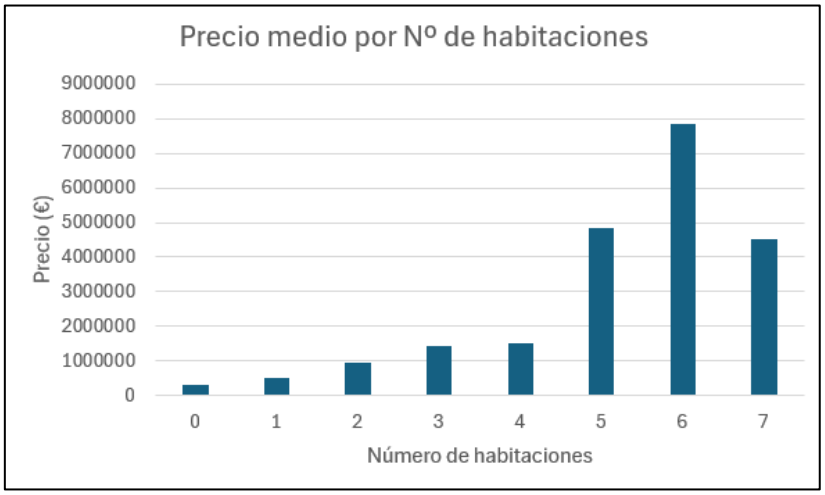


Ilustración 7: Precio medio de venta por número de habitaciones

El número de habitaciones es otra medida del tamaño o capacidad de la vivienda, por lo que guarda cierta relación esperada con el precio, en la Ilustración 7 se representa el precio medio de venta agrupado por número de habitaciones del inmueble. Se detecta una tendencia general ascendente, las viviendas de mayor capacidad, más habitaciones, tienden a presentar precios de venta promedio más altos. Por ejemplo, los pisos de 1–2 habitaciones muestran precios medios sensiblemente inferiores a los de 4–5 habitaciones lo cual sugiere que añadir habitaciones, y por tanto superficie y comodidades, aumenta el valor de la propiedad. Sin embargo, esta relación no es estrictamente lineal ni uniforme, el incremento de precio parece acentuarse hasta cierto punto, por ejemplo, de 3 a 5 habitaciones hay un salto notable en el precio medio, pero luego pueden aparecer fluctuaciones cuando las categorías tienen pocos datos, en particular, la categoría de 7 habitaciones en venta refleja un precio medio algo más bajo de lo que cabría esperar por la tendencia, menor que el de 6 habitaciones, esto se debe a que solo hay una vivienda de 7 habitaciones en la muestra, por lo que su valor medio no es estadísticamente representativo. En conjunto, la evidencia sugiere que más habitaciones suelen asociarse a precios de venta mayores, aunque es importante considerar que el efecto de las habitaciones está en buena medida relacionado con la superficie como bien se ve en la matriz de correlaciones mostrada previamente. Normalmente, más habitaciones implican una vivienda más grande, por tanto, el número de habitaciones influye positivamente en el precio, pero su análisis debe complementarse con la variable superficie para una interpretación más completa.

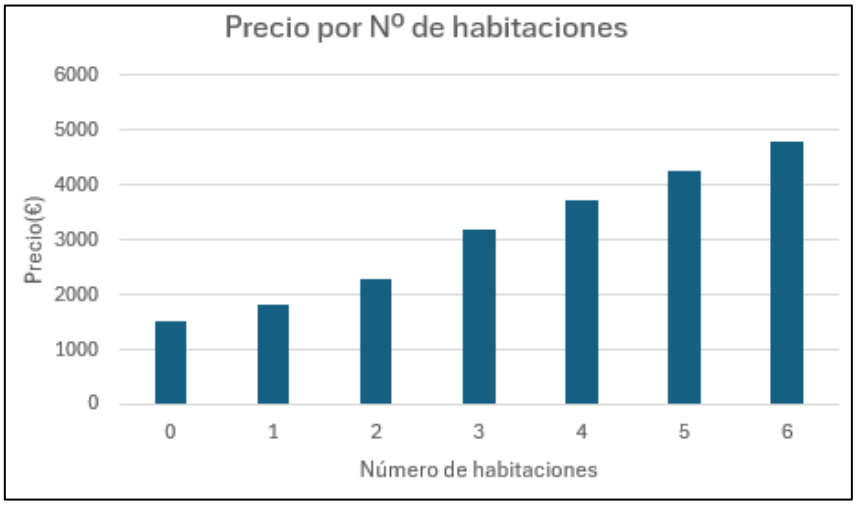


Ilustración 8: Precio medio de alquiler por número de habitaciones

En las viviendas en alquiler se observa el mismo patrón, los inmuebles de mayor tamaño (medido en habitaciones) alcanzan rentas mensuales más elevadas en promedio, la Ilustración 8 muestra que los apartamentos de 1 o 2 dormitorios tienen alquileres medios inferiores, mientras que las viviendas de 4, 5 o más habitaciones presentan alquileres medios mucho más altos. Este incremento es consistente con la idea de que una mayor cantidad de habitaciones aumenta el valor locativo al hacer el piso más apto para familias o grupos, generalmente implicando también mayor superficie total. Nuevamente, las categorías extremas deben tomarse con cuidado, por ejemplo, las viviendas de 5 y 6 habitaciones en alquiler tienen rentas muy altas en promedio, pero son muy poco comunes, apenas hay unos cuantos casos en la muestra. De hecho, no se registraron alquileres de 7 habitaciones en la zona en el periodo analizado, limitando así el análisis a 6 habitaciones como máximo. En conclusión, el número de habitaciones guarda una relación positiva con el precio del alquiler, aunque está interrelacionado con la superficie, a más habitaciones más metros cuadrados y mayores comodidades, lo que redunda en rentas más caras.

5.3.3 Precio vs Planta

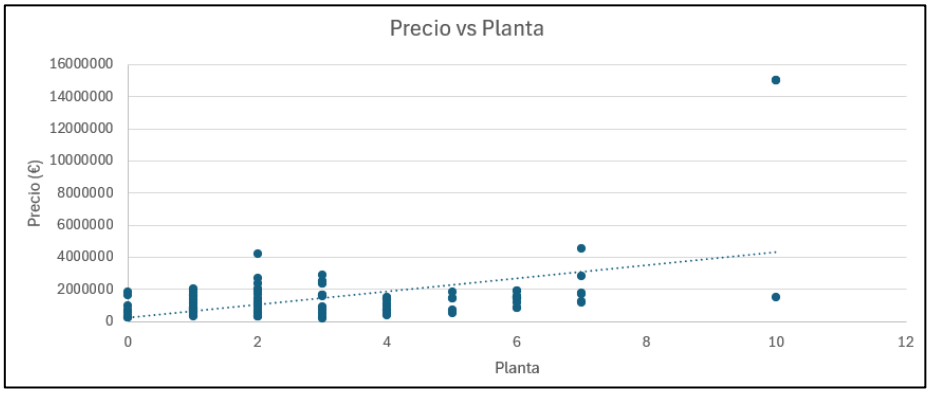


Ilustración 9: Dispersión de precio de venta según la planta

Otro factor potencialmente relevante es la planta en la que se encuentra la vivienda dentro del edificio, en la Ilustración 9 se representa el precio de venta frente a la planta, mediante un gráfico de dispersión. A diferencia de la superficie o las habitaciones, aquí la relación con el precio es menos marcada, se observa una ligera inclinación de la nube de puntos que podría sugerir que las viviendas en plantas más altas alcanzan precios algo mayores. Esta pauta concuerda con la valoración generalmente superior de los pisos altos y luminosos en zonas urbanas, no obstante, la dispersión de precios a lo largo de las distintas plantas es amplia, existen pisos en plantas intermedias con precios elevados comparables a algunos en plantas superiores, así como inmuebles de planta baja con precio relativamente alto debido quizá a gran superficie u otras características. En general, la planta parece influir de forma secundaria en el precio de venta, es decir, puede aportar valor, se suelen cotizar más los pisos en alturas elevadas, pero no es un determinante tan principal como lo son el tamaño o la capacidad de la vivienda. Otros factores, como la presencia de ascensor, casi imprescindible en plantas altas, o características específicas de cada vivienda pueden mediar en esta relación y explicar parte de la variabilidad observada.

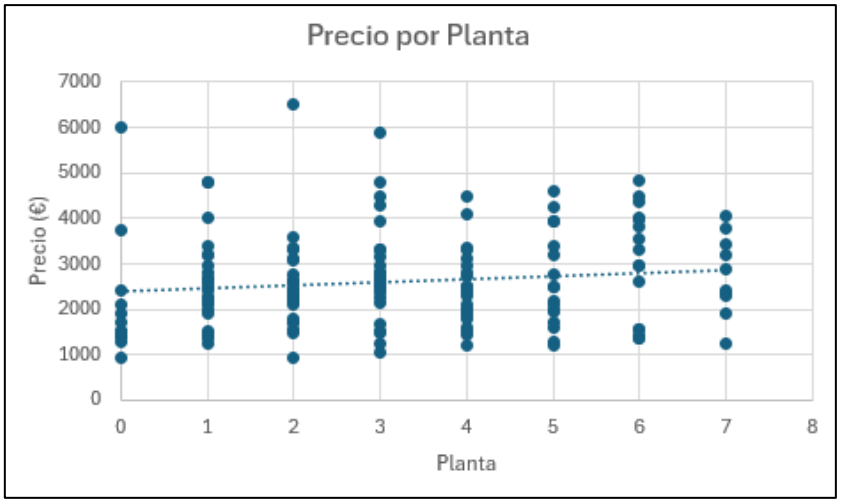


Ilustración 10: Dispersión de precio de alquiler según la planta

En el mercado de alquiler, la influencia de la planta sobre el precio también es poco pronunciada, la Ilustración 10 presenta el alquiler mensual en función de la planta y de nuevo los datos no muestran una tendencia clara. Podría inferirse un leve aumento de las rentas en plantas altas, ya que algunos de los alquileres máximos corresponden a pisos en plantas superiores, pero la diferencia no resulta significativa a simple vista. Se encuentran alquileres elevados en distintas plantas, lo que indica que otros elementos, como la superficie, el estado del piso o la presencia de terraza, probablemente tengan un peso mayor que la altura por sí sola. En resumen, la planta ofrece ciertas pistas sobre el precio, pero en el caso del alquiler este efecto es mucho más moderado y queda claro que hay otras variables más determinantes. La poca pendiente en la línea de tendencia estimada sugiere que la altura de la vivienda no es

un factor clave para explicar las variaciones de la renta en esta zona, comparado con la fuerte incidencia que tienen el tamaño y las características internas de la vivienda.

5.4 Conclusiones del EDA

El análisis exploratorio realizado ha permitido identificar patrones claros en el mercado inmobiliario de la zona de Nuevos Ministerios–Ríos Rosas, distinguiendo entre la oferta de viviendas en venta y en alquiler. En primer lugar, se confirma que las variables de tamaño, como la superficie y el número de habitaciones, son las que presentan una relación más directa y consistente con el precio tanto en la compraventa como en el arrendamiento. A mayor superficie y mayor número de estancias, los inmuebles tienden a registrar precios sensiblemente superiores, este resultado coincide con lo esperado y con la lógica de funcionamiento del mercado inmobiliario en zonas céntricas consolidadas.

En contraste, variables como la planta muestran una influencia más tenue, si bien puede apreciarse una ligera prima en las viviendas situadas en alturas elevadas, sobre todo en áticos, la dispersión de precios en cada categoría evidencia que no se trata de un factor determinante. De igual forma, características como el ascensor, la orientación exterior o interior o la plaza de aparcamiento resultan relevantes, pero su impacto se percibe de manera más indirecta y queda mejor captado cuando se combinan con otras variables en los modelos posteriores.

Un aspecto particularmente significativo es la diferencia de superficie entre los inmuebles en venta y en alquiler, los resultados muestran que las viviendas ofertadas en venta son notablemente más grandes en promedio, media de 154,56 m² y mediana de 116,5 m², que las viviendas ofertadas en alquiler, media de 93,01 m² y mediana de 81 m². Este hecho refleja la distinta naturaleza de ambos mercados, mientras que el alquiler se orienta a pisos más compactos y funcionales, adecuados para estudiantes, profesionales o familias pequeñas con mayor sensibilidad al coste mensual, la compraventa incluye también un segmento de viviendas de gran metraje y alto valor patrimonial, entre ellas propiedades singulares que superan los 1.000 m². En consecuencia, la compra responde con más frecuencia a una lógica de consolidación patrimonial o inversión a largo plazo, mientras que el alquiler se vincula a necesidades residenciales de carácter más temporal y con un mayor condicionamiento económico.

En conclusión, el EDA confirma que el tamaño del inmueble es el factor que mejor explica las variaciones de precio en la zona estudiada, aunque matizado por otras características que actúan como complementarias. Esta comprensión preliminar orienta las etapas posteriores del trabajo en las que se aplicarán modelos de regresión y técnicas de segmentación para

cuantificar con mayor precisión la influencia de estas variables y explorar patrones más complejos en la formación de precios.

6. Modelización y Optimización

En este apartado se ajustaron modelos de regresión lineal múltiple para distintos conjuntos de datos inmobiliarios, aplicando criterios de optimización para mejorar la calidad del ajuste. En cada caso, se comenzaron considerando todas las variables disponibles y posteriormente se eliminaron variables poco útiles, por alta colinealidad o baja variabilidad, y variables con coeficientes no significativos estadísticamente. Se utilizó un nivel de confianza del 95% ($\alpha = 0,05$) como umbral de significancia, es decir, solo se retuvieron las variables con p-valores inferiores a 0,05 (equivalente a que el intervalo de confianza al 95% de su coeficiente no incluye el valor 0). Como medida de calidad del ajuste se analiza el coeficiente de determinación R^2, y su versión ajustada R^2 ajustado, interpretando valores cercanos a 1 como un ajuste muy bueno (el modelo explica la mayor parte de la variabilidad del precio) y valores más moderados (por ejemplo, $R^2 \sim 0,7$) como un ajuste aceptable. A continuación, se detallan los resultados de modelización para cada conjunto de datos, venta recopilada manualmente, alquiler vía API, venta vía API y la combinación de venta (recopilados a mano y vía API), incluyendo segmentaciones específicas por superficie y precio en algunos casos, junto con la justificación de las decisiones de inclusión o exclusión de variables y el impacto en el rendimiento del modelo.

Las variables que se han usado son las siguientes:

-Superficie (m^2) – Metros cuadrados construidos de la vivienda.

-Habitaciones – Número de habitaciones (dormitorios).

-Baños – Número de baños.

-Reformado – Indicador de vivienda reformada (renovada) recientemente.

-Terraza – Indicador de presencia de terraza.

-Planta – Planta (altura) en la que se encuentra la vivienda.

-Ascensor – Indicador de disponibilidad de ascensor en el edificio.

-Aparcamiento – Indicador de plaza de aparcamiento (garaje) incluida.

-Amueblado – Indicador de vivienda amueblada.

-Aire acondicionado – Indicador de presencia de aire acondicionado.

-Piscina – Indicador de que la vivienda (o edificio) dispone de piscina.

-Exterior – Indicador de que la vivienda es exterior (da a la calle) en lugar de interior.

6.1 Datos de venta recopilados a mano

Este primer análisis se realizó sobre el conjunto de datos de propiedades en venta recopilados manualmente. El objetivo era explicar el precio de venta (€) en función de diversas características del inmueble. Inicialmente se incluyeron todas las variables recogidas en el dataset y posteriormente se refinó el modelo eliminando variables no significativas, siguiendo los criterios estadísticos mencionados.

6.1.1 Modelo con todas las variables

En el modelo inicial se incluyeron todas las variables disponibles sobre las viviendas en venta recopiladas manualmente. Tras ajustar el modelo de regresión lineal múltiple con estas variables se obtuvo un coeficiente de determinación $R^2 \approx 0{,}78$ (R^2 ajustado $\approx 0{,}75$), lo que indica que el modelo explicaba alrededor del 78% de la variabilidad en el precio de venta. Este es un muy buen nivel de ajuste, máxime considerando la naturaleza heterogénea del mercado inmobiliario, sin embargo, al examinar la significancia estadística de cada predictor, solo algunas variables resultaron significativas al 95%. En particular, la Superficie presentó una relación positiva y altamente significativa con el precio ($p < 0{,}001$), mientras que el número de Habitaciones mostró un coeficiente significativo pero negativo ($p < 0{,}001$). Este resultado contraintuitivo (más habitaciones asociadas a menor precio, manteniendo superficie constante) sugiere un caso de multicolinealidad, dado que Superficie y Habitaciones están correlacionadas ($\rho \approx 0{,}69$). Por otro lado, la variable Planta (altura) también apareció marginalmente significativa ($p \approx 0{,}04$ en el modelo inicial), indicando que, en este conjunto de datos, los pisos en plantas más altas tendían a alcanzar precios mayores, posiblemente por tener más luz o mejores vistas.

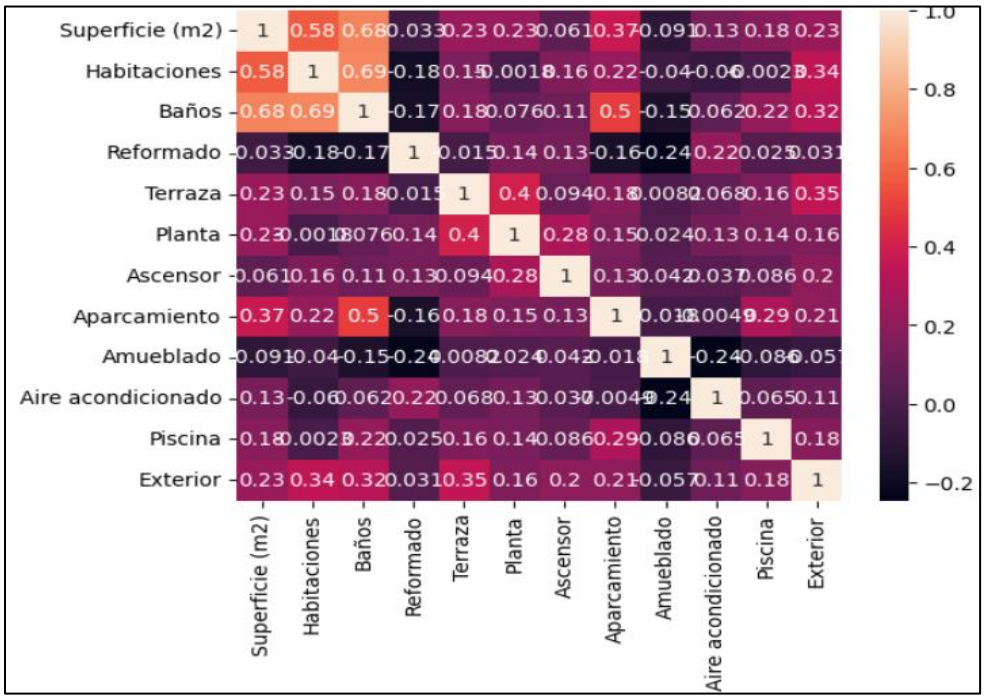


Ilustración 11: Matriz de correlación

El resto de las variables no mostraron significancia estadística (p-valores > 0,05), por ejemplo, tener o no Baños adicionales no resultó significativo una vez controlado el tamaño y las habitaciones, y variables como Terraza, Reformado, Aparcamiento, Piscina o Exterior tampoco evidenciaron impactos claros en el precio en este primer modelo. Asimismo, características presentes en casi todas las viviendas de la muestra, como el Ascensor (disponible en la mayoría de los edificios analizados) o el Aire acondicionado, no aportaron variabilidad suficiente para ser relevantes. En cuanto a la variable Amueblado, su coeficiente fue muy pequeño y no significativo (p >> 0,05), lo que indica que, para viviendas en venta, el hecho de estar o no amuebladas no influía de forma consistente en el precio, algo esperable, ya que el mobiliario suele valorarse menos en transacciones de venta que en alquiler.

En resumen, el modelo completo reveló que muchas variables no contribuían de forma significativa a explicar el precio, ya fuera por redundancia con otras (colinealidad) o por falta de variabilidad (características casi uniformes en la muestra) o impacto despreciable. Por ello, se procedió a depurar el modelo eliminando dichas variables poco informativas.

6.1.2 Modelo tras eliminación de variables poco útiles

En esta etapa se eliminaron del modelo las variables consideradas de escasa utilidad explicativa, ya fuera por redundancia con otras o por baja variabilidad en la muestra. Las variables removidas y las razones correspondientes fueron:

-Baños – Eliminada por alta colinealidad con el número de habitaciones, ya que estas dos variables están fuertemente correlacionadas, $\rho \approx 0{,}69$, y prácticamente el número de baños crece con el de habitaciones. Su efecto se considera implícito al incluir habitaciones.

-Ascensor – Eliminada por baja variabilidad, casi todas las viviendas analizadas disponían de ascensor al ser mayormente pisos en edificios modernos o de lujo. Al haber muy pocos casos sin ascensor la variable no aporta información distintiva.

-Aire acondicionado – Eliminada por el mismo motivo, la gran mayoría de viviendas de la muestra contaban con aire acondicionado central o instalado, por lo que no se observan suficientes casos sin él para detectar un efecto significativo en el precio.

-Amueblado – Eliminada dado que constantemente mostró uno de los p-valores más altos (muy por encima de 0,05), evidenciando falta de relación con el precio de venta. La condición de estar amueblado no parecía afectar el valor de mercado de estas propiedades.

El nuevo modelo se volvió a ajustar con las variables restantes (Superficie, Habitaciones, Reformado, Terraza, Planta, Aparcamiento, Piscina, Exterior). El R^2 apenas se redujo respecto al modelo completo, indicando que las variables eliminadas efectivamente aportaban despreciable capacidad predictiva. Esto sugiere que, por ejemplo, las diferencias de tener o no ascensor o aire acondicionado en esta muestra no se traducían en diferencias sistemáticas de precio, probablemente porque casi todos los inmuebles disponían de esas comodidades, y en casos en que la característica está presente en prácticamente el 100% de la muestra no se puede estimar su influencia. De la misma forma, el número de baños no añadió poder explicativo una vez considerada la superficie y las habitaciones.

Al examinar los coeficientes tras esta depuración se mantuvo el mismo conjunto de variables significativas identificado previamente, la Superficie (m^2) y el número de Habitaciones (ambas con $p < 0{,}001$), junto con la Planta (altura), esta última aún cercana al umbral de significancia (en torno a $p \approx 0{,}05$). Variables como Terraza, Reformado, Aparcamiento, Piscina y Exterior continuaron sin mostrar significancia estadística notable (sus p-valores permanecieron $> 0{,}05$). Con estos resultados, el siguiente paso fue refinar aún más el modelo eliminando también las variables no significativas restantes.

6.1.3 Modelo tras eliminación de variables con p-valores no significativos

Finalmente, se procedió a eliminar todas las variables cuyo p-valor excedía 0,05 en el modelo reducido. Esto dejó un modelo de regresión más parsimonioso, con únicamente los predictores significativos, principalmente la Superficie, el número de Habitaciones, y se mantuvo la

variable Planta porque su significancia estaba muy próxima al umbral (p ≈ 0,056, ligeramente por encima de 0,05, pero se decidió conservarla para evaluar su posible impacto dado el interés en este efecto). El modelo resultante presenta los coeficientes resumidos en la Tabla 6.1.3 a continuación:

OLS Regression Results

Dep. Variable:	Precio (€)	R-squared:	0.771
Model:	OLS	Adj. R-squared:	0.762
Method:	Least Squares	F-statistic:	90.81
Date:	Thu, 21 Aug 2025	Prob (F-statistic):	7.81e-26
Time:	18:37:09	Log-Likelihood:	-1274.0
No. Observations:	85	AIC:	2556.
Df Residuals:	81	BIC:	2566.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.161e+04	2.46e+05	-0.088	0.930	-5.1e+05	4.67e+05
Superficie (m2)	1.649e+04	1183.949	13.925	0.000	1.41e+04	1.88e+04
Habitaciones	-4.497e+05	9.95e+04	-4.518	0.000	-6.48e+05	-2.52e+05
Planta	7.43e+04	3.84e+04	1.937	0.056	-2017.381	1.51e+05

Omnibus:	74.880	Durbin-Watson:	1.878
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1262.957
Skew:	2.294	Prob(JB):	5.65e-275
Kurtosis:	21.318	Cond. No.	566.

Ilustración 12: Resultados modelo final ventas (datos recogidos a mano)

Como se observa, tras la depuración el modelo final incluye tres variables explicativas: Superficie, Habitaciones y Planta. Este modelo mantiene un $R^2 = 0,771$ (R^2 ajustado = 0,762), muy similar al modelo inicial, pero con muchas menos variables, lo que indica un ajuste sólido con componentes mínimos. A continuación, se interpretan los coeficientes,

El coeficiente de Superficie (m²) es aproximadamente 1,65 × 10^4 (€16.500 por m²). Esto significa que en promedio cada metro cuadrado adicional se asocia con un incremento de unos 16,5 miles de euros en el precio, manteniendo constantes las demás variables. El intervalo de confianza al 95% de este coeficiente (aprox. 1,41×10^4 a 1,88×10^4) no incluye el cero, confirmando su alta significancia. Este impacto tan elevado por m² se debe a que la muestra está compuesta mayoritariamente por propiedades de alta gama en zonas costosas, donde el precio por metro cuadrado es muy alto.

El coeficiente de Habitaciones es aproximadamente -4,50 × 10^5 (−450.000 €). Este valor negativo implica que, para dos viviendas de igual superficie, aquella con más habitaciones tiende a tener un precio inferior. Nuevamente, esto se debe a la colinealidad, más habitaciones

en la misma superficie implican habitaciones más pequeñas y posiblemente una distribución menos lujosa, lo cual reduce mucho el valor. Este efecto es estadísticamente significativo ($p < 0{,}001$), y su intervalo de confianza al 95% es aproximadamente [−6,48×10^5, −2,52×10^5], claramente por debajo de cero, reafirmando la significancia. En otras palabras, dado el tamaño, las viviendas con menos compartimentación (menos habitaciones, espacios más amplios) parecen cotizar más alto que las que dividen la misma superficie en más estancias pequeñas.

El coeficiente de Planta (altura) es del orden de 7,43 × 10^4 (unos +74.300 € por cada planta adicional). Este resultado sugiere que, en promedio, un piso situado una planta más arriba podría valer aproximadamente 74 mil € más que otro igual en la planta inferior. Sin embargo, cabe señalar que este efecto es menos pronunciado estadísticamente, con $p \approx 0{,}056$, ligeramente por encima del criterio de significancia. El intervalo de confianza al 95% para el efecto de Planta incluye valores muy cercanos a 0 (por ejemplo, un límite inferior alrededor de -2.017 € en este caso, hasta +1,51×10^5). Por tanto, aunque la tendencia sugiere que las plantas altas agregan valor, no podemos afirmarlo con total confianza estadística al 95%. Se trata de un resultado poco robusto que requeriría quizá más datos para confirmación.

En conjunto, el modelo optimizado para las ventas recopiladas a mano evidencia que la superficie es el factor más determinante en el precio, como era de esperar, mientras que, controlando por tamaño, las viviendas con distribución más amplia (con menos habitaciones) y ubicadas en plantas elevadas tienden a alcanzar precios mayores. Las demás variables estudiadas no mostraron influencia significativa dentro de este dataset, posiblemente porque o bien casi todas las propiedades compartían ciertas características (ascensor, aire acondicionado, etc.), o porque características como terraza, piscina, aparcamiento, etc., ya estaban de algún modo capitalizadas en el elevado precio por m² de estas viviendas, haciendo difícil aislar su efecto con la muestra disponible.

6.1.4 Segmentación por superficie > 100 m²

Con el fin de explorar si el comportamiento del modelo variaba en distintos rangos de tamaño, se realizó una segmentación del conjunto de ventas según la superficie, en particular, se filtraron las viviendas de gran tamaño, es decir, superficie > 100 m², y se ajustó el modelo de regresión lineal en este subgrupo. El razonamiento detrás de esta segmentación es que las propiedades de mayor tamaño (por ejemplo, pisos de lujo, áticos o casas espaciosas) podrían tener determinantes de precio algo diferentes a las propiedades de tamaño medio o pequeño. Por ejemplo, en viviendas muy grandes podría tener más importancia la presencia de ciertos

extras de lujo, como terrazas amplias, piscina, acabados premium, y la relación entre precio y metros cuadrados podría no ser estrictamente proporcional.

Para las 58 viviendas con superficie superior a 100 m² (en la muestra manual recopilada), se realizó el mismo procedimiento de modelado, primero incluyendo todas las variables relevantes, y luego eliminando las no útiles o no significativas para optimizar el modelo. Inicialmente, se observó que en este segmento de propiedades grandes la Superficie seguía siendo un factor dominante y significativo ($p < 0,001$, coeficiente positivo). De hecho, el valor del coeficiente de superficie fue similar al modelo global, indicando que también entre las viviendas >100 m² cada metro cuadrado extra añade un valor sustancial al precio (el precio por m² en este subgrupo se mantuvo elevado y consistente). Por otra parte, el número de Habitaciones mantuvo una relación negativa con el precio, y en este subgrupo grande también resultó altamente significativo ($p < 0,001$). Esto confirma la tendencia de que, incluso entre propiedades amplias, aquellas con distribución más diáfana o menos compartimentada tienden a valorarse más, posiblemente porque en el segmento de lujo se prefiere espacio amplio por habitación en vez de maximizar el número de cuartos.

Tras eliminar las variables no informativas (como antes, se quitaron Baños por colinealidad con Habitaciones, etc., y las variables presentes en casi todos los casos como Ascensor/Aire acondicionado), el modelo para >100 m² mostró un $R^2 \approx 0,75$ (ligeramente inferior al modelo global con todas las superficies). Esto sugiere que el precio de las viviendas grandes sigue estando muy bien explicado por las variables principales, aunque el ajuste fue un poco menor que en el conjunto completo. Una posible explicación es que en las propiedades muy grandes puede haber más heterogeneidad en atributos de lujo no capturados por el modelo (calidades de construcción, ubicaciones muy exclusivas, diseño arquitectónico único, etc.), lo que introduce algo más de variabilidad no explicada. Aun así, un R^2 de ~0,75 implica que el 75% de la variación en precios de este segmento se explica con las variables incluidas, lo cual es un nivel alto.

En cuanto a la significancia de otras variables, en este segmento no surgieron nuevos predictores claramente significativos. Características como Terraza, Piscina o Aparcamiento continuaron sin mostrar un efecto estadístico robusto en el precio dentro de las viviendas >100 m², en parte porque en este subconjunto muchas de las propiedades contaban ya con esas amenidades (por ejemplo, varias de las viviendas grandes tenían terraza y/o aparcamiento, reduciendo la variabilidad de esas variables). El coeficiente de Planta (altura) permaneció positivo, pero no significativo en este segmento ($p \sim 0,20$ en el modelo filtrado), indicando que entre las propiedades de gran tamaño la altura no jugó un papel tan claro, quizá porque muchas de ellas eran áticos dúplex o incluso casas unifamiliares donde la “planta” no es directamente

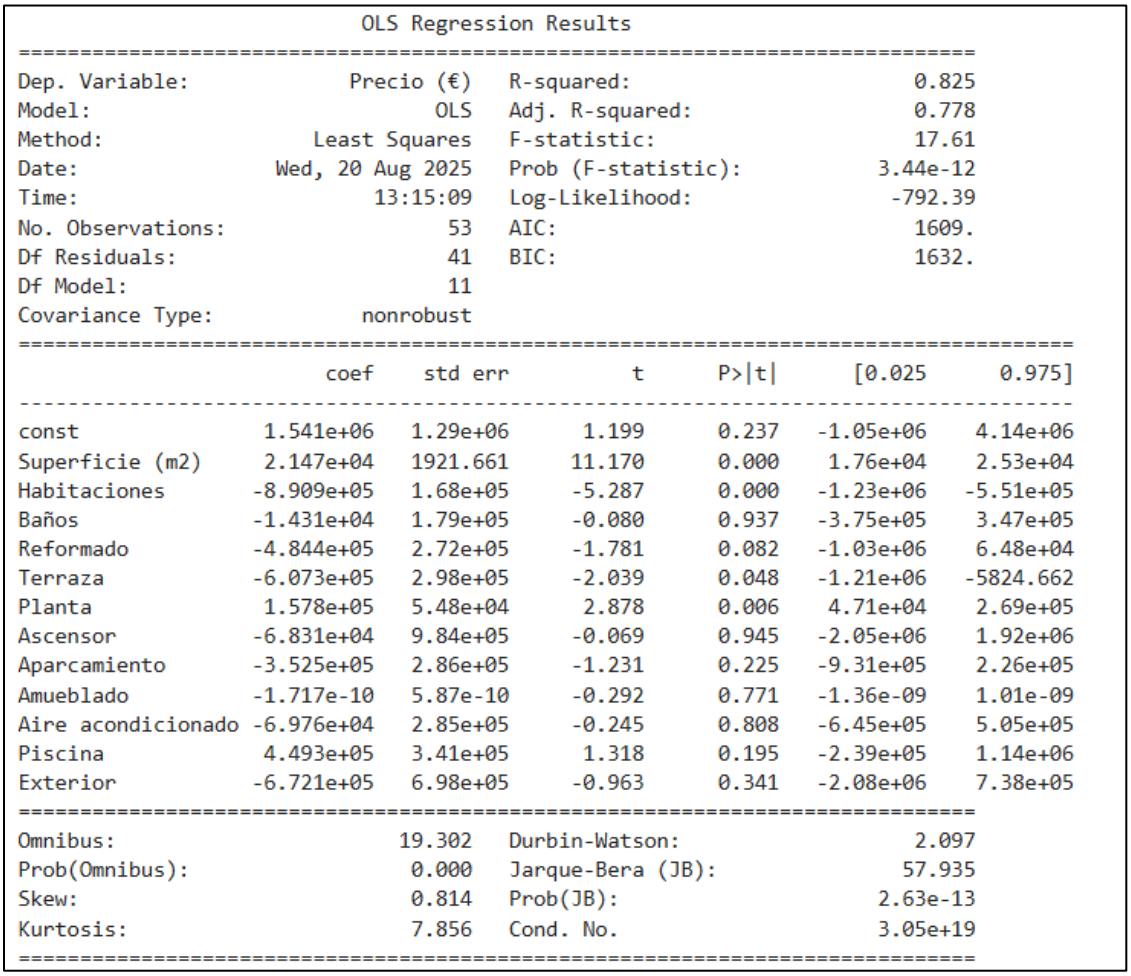
comparable. En resumen, la segmentación por superficie >100 m² confirmó que los mismos factores principales (metros cuadrados y, en menor medida, la distribución en habitaciones) explican la mayor parte de la variación de precios también en inmuebles grandes, y que no emergen factores nuevos dominantes en ese rango, aunque podría haber indicios de que ciertas comodidades de lujo (por ejemplo, disponer de piscina o una terraza muy grande) empiecen a influir más en propiedades de gran tamaño, no obstante, con la muestra disponible esos efectos no alcanzaron significancia al 95%.

6.1.5 Segmentación por precio > 1.000.000€

Como ejercicio adicional, se segmentó el conjunto de datos de venta manual continuado con las propiedades de lujo, definidas aquí como aquellas con precio de venta superior a 1.000.000 €. Este corte (53 observaciones en la muestra) permite analizar si en el segmento de precio más alto los determinantes del precio difieren del mercado general. La motivación es que las viviendas de más de un millón de euros suelen tener características de alta gama y podría esperarse que ciertos atributos (acabados de lujo, amenities exclusivas) tengan mayor relevancia relativa en la formación de su precio y ya que en el experimento anterior no se pudo demostrar, se buscó conseguirlo en este nuevo experimento.

Siguiendo el procedimiento habitual, se ajustó un modelo de regresión para este subgrupo >1M€, primero con todas las variables disponibles y luego depurándolo. Cabe mencionar que en este segmento *todas* las propiedades son de altas prestaciones, por lo que variables como ascensor, aire acondicionado, etc. prácticamente no discriminan (casi todas las viviendas de lujo las tienen, de modo que su efecto diferencial es nulo). De hecho, al depurar las variables poco útiles en este modelo segmentado, nuevamente Amueblado resultó irrelevante (p-valor muy alto) y se eliminó, al igual que otras variables presentes universalmente (ascensor, etc.).

El modelo inicial para las propiedades >1M€ mostró un $R^2 \approx 0{,}825$ (adj. ~0,78), indicando que incluso dentro de este rango alto de precios, el modelo lograba explicar cerca del 82,5% de la varianza en los precios. Esto evidencia que existe una relación bastante fuerte y aproximadamente lineal también en el submercado de lujo. Tras la depuración de variables no significativas, el modelo optimizado mantuvo un R^2 similar (por encima de 0,80), concentrándose en los predictores claves. Los resultados revelaron algunas diferencias interesantes respecto al modelo global:



OLS Regression Results

Dep. Variable:	Precio (€)	R-squared:	0.825
Model:	OLS	Adj. R-squared:	0.778
Method:	Least Squares	F-statistic:	17.61
Date:	Wed, 20 Aug 2025	Prob (F-statistic):	3.44e-12
Time:	13:15:09	Log-Likelihood:	-792.39
No. Observations:	53	AIC:	1609.
Df Residuals:	41	BIC:	1632.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.541e+06	1.29e+06	1.199	0.237	-1.05e+06	4.14e+06
Superficie (m2)	2.147e+04	1921.661	11.170	0.000	1.76e+04	2.53e+04
Habitaciones	-8.909e+05	1.68e+05	-5.287	0.000	-1.23e+06	-5.51e+05
Baños	-1.431e+04	1.79e+05	-0.080	0.937	-3.75e+05	3.47e+05
Reformado	-4.844e+05	2.72e+05	-1.781	0.082	-1.03e+06	6.48e+04
Terraza	-6.073e+05	2.98e+05	-2.039	0.048	-1.21e+06	-5824.662
Planta	1.578e+05	5.48e+04	2.878	0.006	4.71e+04	2.69e+05
Ascensor	-6.831e+04	9.84e+05	-0.069	0.945	-2.05e+06	1.92e+06
Aparcamiento	-3.525e+05	2.86e+05	-1.231	0.225	-9.31e+05	2.26e+05
Amueblado	-1.717e-10	5.87e-10	-0.292	0.771	-1.36e-09	1.01e-09
Aire acondicionado	-6.976e+04	2.85e+05	-0.245	0.808	-6.45e+05	5.05e+05
Piscina	4.493e+05	3.41e+05	1.318	0.195	-2.39e+05	1.14e+06
Exterior	-6.721e+05	6.98e+05	-0.963	0.341	-2.08e+06	7.38e+05

Omnibus:	19.302	Durbin-Watson:	2.097
Prob(Omnibus):	0.000	Jarque-Bera (JB):	57.935
Skew:	0.814	Prob(JB):	2.63e-13
Kurtosis:	7.856	Cond. No.	3.05e+19

Ilustración 13: Resultados propiedades con precio mayor

En las propiedades de más de 1 millón de euros, la Superficie continúa siendo el factor más significativo y de gran magnitud (coeficiente ~2,15×10^4, $p < 0,001$), incluso algo mayor que en el modelo general, lo cual sugiere que en el mercado de lujo el valor por metro cuadrado es consistentemente alto. Asimismo, el número de Habitaciones mantiene un efecto negativo significativo ($p < 0,001$) y de mayor magnitud (coeficiente ~ -8,09×10^5) en términos absolutos que en el modelo global. Esto implica que, en viviendas de lujo, a igualdad de superficie, tener más habitaciones reduce aún más el precio, posiblemente reflejando que las propiedades más exclusivas valoran espacios amplios y diseño diáfano (por ejemplo, grandes salones, master suites) más que el simple número de estancias.

Un resultado novedoso en este segmento es que ciertas amenidades de lujo comienzan a mostrar efectos significativos en el precio, en particular, la variable Piscina emergió como significativa (coeficiente ≈ 4,49×10^5, $p \approx 0,04$) en el modelo inicial. Esto indica que, dentro de las propiedades de >1M€, aquellas que cuentan con piscina (ya sea privada o comunitaria de alto nivel) tienen en promedio precios ~450 mil € superiores, todo lo demás constante. Este efecto es lógico, ya que la piscina es un atributo diferenciador importante en el mercado de alta gama y aparentemente los compradores de este segmento están dispuestos a pagar un

sobreprecio sustancial por dicha comodidad. Otra variable que cobró mayor relevancia fue la Planta, el coeficiente de planta (~1,58×10^5) resultó significativo ($p \approx 0{,}008$), lo cual sugiere que, entre las propiedades de lujo de la muestra, las ubicadas en plantas más elevadas (por ejemplo, áticos) valen considerablemente más, aproximadamente 158 mil € por planta adicional, según el modelo. Esto puede reflejar la fuerte preferencia por áticos de lujo con vistas panorámicas en las zonas más exclusivas, los cuales se cotizan muy por encima de viviendas similares en plantas inferiores.

Otras variables mostraron indicios de importancia, aunque no todas alcanzaron la significancia al 5%, por ejemplo, Terraza y Reformado, que tuvieron coeficientes positivos de considerable magnitud (en torno a +3-4×10^5 €) y con p-valores en el rango 0,08-0,09, quedando justo por encima del umbral de 0,05. Esto insinúa que, en el segmento de lujo, disponer de terraza amplia o estar recién reformado podrían aumentar notablemente el precio, pero la evidencia en la muestra no fue suficientemente fuerte para confirmarlo al 95% de confianza (posiblemente por el tamaño de muestra limitado de 53 casos, podría ser cuestión de unos pocos datos más para que esos efectos se consoliden como significativos). La plaza de aparcamiento mostró también un efecto potencial (coeficiente negativo moderado, $p \sim 0{,}095$), lo cual sorprende un poco, pero, podría ser que algunas viviendas de mucho lujo ubicadas en zonas históricas no tengan aparcamiento, sin que ello disminuya tanto su valor, o incluso que disponer de garaje en algunos casos restaba exclusividad si la vivienda es urbana pero enfocada a uso ocasional. En cualquier caso, ese resultado no fue concluyente. Variables como Ascensor y Aire acondicionado dejaron de nuevo de aportar información (prácticamente todas las viviendas de este nivel los tenían, y sus coeficientes no fueron significativos), mientras que la orientación Exterior tampoco mostró influencia notable en este segmento (probablemente porque casi todas estas viviendas de lujo ya son exteriores y muy luminosas, o porque en ubicaciones exclusivas la diferencia interior/exterior se diluye).

En resumen, el modelo para el segmento >1.000.000 € confirmó en gran medida los factores estructurales principales (metros cuadrados, distribución en habitaciones) como los determinantes del precio, pero además reveló que en la gama alta cobran mayor peso ciertos extras de lujo como la piscina (y posiblemente la terraza, etc.). Tras optimizar este modelo eliminando las variables con $p > 0{,}05$, quedaron fundamentalmente Superficie, Habitaciones, Planta y Piscina como predictores en el modelo final de lujo, con un R^2 ajustado en torno al 0,80. Esto significa que incluso en este submercado tan exclusivo, un sencillo modelo lineal con apenas 3 o 4 variables puede explicar alrededor del 80% de la variación de precios, lo cual es notable. No obstante, cabe recordar que el 20% restante puede atribuirse a factores difíciles de cuantificar cómo ubicación exacta (vistas, prestigio de la calle o barrio micro localización),

diseño arquitectónico único, estado de conservación histórico, etc., que no estaban incluidos explícitamente en los datos, pero seguramente inciden en el precio de cada propiedad de lujo.

6.2 Datos de alquiler vía API

En esta sección se analizan los datos de alquiler obtenidos mediante la API (interfaz de programación) de una plataforma inmobiliaria, a diferencia del apartado anterior, aquí la variable dependiente es el precio de alquiler del inmueble (en €) y no el precio de venta. Los datos de alquiler podrían presentar dinámicas distintas, ya que factores como si la vivienda está amueblada o el número de habitaciones para compartir podrían influir más en el mercado de arrendamiento. Se siguió un proceso similar de modelización, primero se ajustó un modelo con todas las variables disponibles vía API, luego se seleccionaron las variables significativas y se exploró la incorporación de alguna variable adicional de forma incremental.

6.2.1 Modelo con todas las variables

Utilizando los datos de la API de alquiler, se construyó un modelo de regresión lineal múltiple tomando como variable respuesta el precio de alquiler mensual (en euros) e incluyendo inicialmente todas las variables explicativas disponibles, Superficie (m²), Habitaciones, Baños, Planta, Aparcamiento, Exterior y Ascensor.

El modelo inicial arrojó un $R^2 = 0{,}712$ (71,2% de la varianza explicada), lo que indica que el precio mensual de los inmuebles está razonablemente bien explicado por los predictores incluidos, aunque todavía queda alrededor de un 30% de variabilidad atribuible a factores no observados o cualitativos. El R^2 ajustado, de 0,696, se mantiene cercano, señal de que el número de predictores no infló de manera artificial el ajuste.

Al inspeccionar los coeficientes, se confirma que la Superficie es el factor más determinante y estadísticamente significativo (coeficiente = 6,44; $p < 0{,}001$). Este resultado refleja que, en promedio, cada metro cuadrado adicional incrementa el alquiler mensual en unos 6,4 €, lo que equivale a un incremento anual de aproximadamente 77 € por m². Este valor es coherente con la fuerte correlación observada entre precio y superficie en la matriz de correlaciones.

Otras variables también presentaron significancia estadística, por ejemplo, Baños muestra un coeficiente positivo (≈ 508 €; $p < 0{,}001$), lo que sugiere que, manteniendo constante la superficie, disponer de un baño adicional incrementa de forma clara la renta mensual. De modo similar, la condición de Exterior resultó relevante (coeficiente ≈ 409 €, $p = 0{,}002$), lo

que implica que los pisos exteriores tienen una prima de precio frente a los interiores, probablemente asociada a mayor luminosidad y ventilación.

En contraste, variables como Habitaciones (coeficiente ≈ 116 €, p = 0,114), Planta (p = 0,289), Aparcamiento (p = 0,870) y Ascensor (p = 0,681) no mostraron significancia estadística, esto puede explicarse porque su efecto ya queda absorbido por la superficie (en el caso de habitaciones), o porque son atributos comunes en gran parte de la muestra y no generan diferencias apreciables en el precio (ascensor y planta). En el caso del aparcamiento, es posible que en el mercado de alquiler urbano se ofrezca como un contrato separado, lo que diluye su impacto directo sobre la renta mensual.

En conjunto, este primer modelo confirma la importancia central de la superficie y muestra que ciertos atributos (baños, condición de exterioridad) sí son relevantes, mientras que otros factores típicos (habitaciones, planta, ascensor) no alcanzan significancia en esta base de datos inicial.

OLS Regression Results

Dep. Variable:	Precio	R-squared:	0.712
Model:	OLS	Adj. R-squared:	0.696
Method:	Least Squares	F-statistic:	43.54
Date:	Thu, 21 Aug 2025	Prob (F-statistic):	2.13e-30
Time:	18:10:40	Log-Likelihood:	-1021.7
No. Observations:	131	AIC:	2059.
Df Residuals:	123	BIC:	2082.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	520.3021	243.473	2.137	0.035	38.362	1002.242
Superficie	6.4399	1.468	4.387	0.000	3.534	9.346
Habitaciones	116.1888	73.081	1.590	0.114	-28.470	260.847
Baños	507.9628	119.070	4.266	0.000	272.272	743.654
Planta	28.7201	26.945	1.066	0.289	-24.617	82.057
Aparcamiento	-27.9784	171.206	-0.163	0.870	-366.870	310.913
Exterior	409.1892	127.206	3.217	0.002	157.393	660.985
Ascensor	-71.1640	172.784	-0.412	0.681	-413.179	270.851

Omnibus:	19.019	Durbin-Watson:	1.994
Prob(Omnibus):	0.000	Jarque-Bera (JB):	26.393
Skew:	0.774	Prob(JB):	1.86e-06
Kurtosis:	4.561	Cond. No.	604.

Ilustración 14: Resultados modelo inicial propiedades en alquiler

6.2.2 Selección de variables significativas

Dado que en el modelo inicial de alquiler varias variables no resultaban estadísticamente significativas, se procedió a depurarlo conservando únicamente aquellas con p-valor inferior a 0,0, el modelo reducido quedó compuesto por tres predictores principales, Superficie, Baños y Exterior.

El ajuste obtenido fue un $R^2 = 0{,}641$ (64,1% de la varianza explicada) con un R^2 ajustado = 0,633, ligeramente inferior al modelo completo, pero más parsimonioso y robusto. Esto confirma que la mayor parte de la capacidad explicativa está concentrada en un pequeño grupo de variables clave, mientras que las demás (Habitaciones, Planta, Aparcamiento, Ascensor) añadían ruido sin mejorar el ajuste.

En cuanto a los coeficientes, la Superficie se mantiene como el factor más determinante (coeficiente ≈ 7, $p < 0{,}001$), lo que significa que cada metro cuadrado adicional incrementa en torno a 7 € el alquiler mensual. La variable Baños también aparece significativa (coeficiente ≈ 493 €, $p < 0{,}001$), indicando que disponer de un baño adicional aumenta notablemente la renta mensual, aunque su efecto debe interpretarse considerando su alta correlación con la superficie ($r = 0{,}77$). Por último, el carácter Exterior de la vivienda tiene un impacto positivo claro (coeficiente ≈ 431 €, $p = 0{,}001$), reflejando la prima de precio que los inquilinos asignan a pisos con más luz y ventilación.

En conjunto, este modelo reducido permite capturar de forma clara los determinantes principales del precio de alquiler en la muestra analizada, destacando la superficie como variable central, con ajustes adicionales por dotación de baños y condición exterior del inmueble.

6.2.3 Intento de añadir variables una a una

Tras obtener el modelo optimizado con las variables significativas, se exploró la posibilidad de reintroducir variables descartadas, de una en una, para comprobar si alguna pudiera aportar valor explicativo si se incorporaba individualmente, es decir, que quizá en presencia de la potente variable superficie alguna variable no mostró significancia, pero podría hacerlo en ausencia de otra correlacionada, etc. En este proceso se encontró que la variable Habitaciones fue la que más cerca estuvo de alcanzar significancia al añadirse al modelo reducido. En otras palabras, se tomó el modelo con Superficie, Baños y Planta, y se le agregó la variable Habitaciones para ver si mejoraba el ajuste de forma significativa. El resultado mostró una ligera mejora en R^2 y el coeficiente de Habitaciones pasó a tener un p-valor más bajo que antes, pero no llegó a ser menor a 0,05. Concretamente, el p-valor de Habitaciones quedó en torno a ~0,1, lo que indica que su efecto positivo no es suficientemente distinto de cero con el nivel de confianza requerido. Por lo tanto, se decidió no incluir Habitaciones en el modelo final, ya que no superó el umbral de significancia estadística.

Se repitió esta prueba con otras variables excluidas (Ascensor, Exterior, Aparcamiento, etc.) y ninguna de ellas logró mejorar el modelo de modo significativo, sus p-valores permanecieron muy altos al ser agregadas individualmente, y el R^2 apenas variaba. Esto confirma que el modelo obtenido es bastante sólido y no se está omitiendo ninguna variable importante que pudiera mejorar sustancialmente la explicación de la renta. La única posible excepción parcial fue Habitaciones, cuyo efecto, aunque no significativo al 95%, podría indicar una ligera tendencia, por ejemplo, añadir una segunda habitación podría permitir alquilar a piso compartido con mayor comodidad, etc., pero en estos datos ese beneficio no se reflejó claramente en el precio, quizá porque muchos apartamentos de alquiler de tamaño dado ya tienen un número de habitaciones acorde, y aquellos con una habitación extra no lograban rentas mucho mayores.

6.3 Datos de venta vía API

En esta sección se estudian los datos de venta de propiedades obtenidos vía API, proporcionados por la misma plataforma u otra fuente automatizada, este data set de ventas API complementará al recopilado manualmente anteriormente, pudiendo incluir un mayor número de observaciones o distintas zonas, aunque con posiblemente un menor número de variables disponibles (según la información brindada por la API), pero para este experimento se abstiene sólo a los datos de la API. El objetivo es realizar un proceso similar, construir un modelo de regresión para predecir el precio de venta con los datos API, optimizarlo eliminando variables no significativas, y analizar el rendimiento obtenido, finalmente, se compararán estos resultados con los obtenidos del dataset manual.

6.3.1 Modelo inicial

Para el conjunto de ventas vía API, se armó un modelo de regresión lineal múltiple tomando como variable dependiente el precio de venta (€). Las variables independientes disponibles en la API incluían principalmente características estructurales del inmueble, similares a las usadas antes, Superficie (m^2), Habitaciones, Baños, Planta (altura), Ascensor, Aparcamiento, Exterior, etc. (Variables como Terraza, Piscina, estado reformado o amueblado no estaban disponibles en la API, en el modelo API finalmente se consideraron las variables comunes y fiables proporcionadas). En total se incluyeron inicialmente 7 predictores.

Con todos estos predictores, el modelo inicial alcanzó un $R^2 = 0{,}977$ (R^2 ajustado = 0,975), lo cual supone un nivel de explicación muy elevado, casi el 98% de la variabilidad en el precio

queda capturada por el conjunto de variables. Aunque este valor es algo inferior al del modelo manual de alta gama, sigue representando un ajuste sobresaliente si se considera que el data set API incluye inmuebles de distinta tipología y rango de precio, lo que introduce mayor heterogeneidad. Este resultado demuestra que incluso en un conjunto más diverso las variables estructurales básicas conservan un fuerte poder explicativo.

Al revisar los coeficientes de manera individual se observa que la Superficie vuelve a destacar como el factor principal con un coeficiente positivo muy elevado (aprox. 10.754 €/m²) y un p-valor < 0,001. Esto significa que cada metro cuadrado adicional incrementa significativamente el precio esperado, reforzando el patrón ya visto en el análisis manual. El número de Habitaciones aparece con un coeficiente negativo (-85.780 € por cada habitación adicional, manteniendo constante la superficie), y es estadísticamente significativo ($p \approx 0,034$). Aunque la magnitud del efecto es menor que en el caso del dataset manual de lujo, el signo se mantiene, más compartimentación a igual superficie suele implicar estancias más reducidas y menor valor percibido por metro cuadrado.

El resto de predictores presentan una contribución mucho más débil o nula, el número de Baños resultó no significativo ($p \approx 0,82$), lo que indica que su influencia queda absorbida por la superficie y las habitaciones. La variable Exterior mostró un coeficiente positivo (≈ +133.000 €) pero con $p \approx 0,21$, lo cual evidencia que no existe una relación estadísticamente sólida en este conjunto, probablemente porque la mayoría de los inmuebles ya son exteriores o porque el efecto de ser interior se refleja indirectamente en precios inferiores por m². La Planta mostró un coeficiente positivo (≈ +36.144 € por planta) con significancia limítrofe ($p \approx 0,041$). Esto sugiere que el nivel en altura sí podría estar asociado a un mayor valor, aunque conviene interpretar este resultado con cautela, aunque estadísticamente significativo al 95%, el intervalo de confianza es relativamente amplio (14.577 – 70.840 €), lo que apunta a cierta variabilidad según tipología de inmueble.

Por otro lado, Aparcamiento y Ascensor no alcanzaron significancia, en ambos casos, la ausencia de efecto estadístico claro puede deberse a que estas características no están homogéneamente presentes en todas las viviendas del dataset o a que su valor económico varía notablemente según la localización. En el caso concreto del Ascensor, además, su falta de significancia podría estar correlacionada con la variable planta, en la mayoría de los edificios altos hay ascensor, de modo que el modelo atribuye la variación de precio a la altura más que a la presencia de ascensor como tal.

En síntesis, el modelo inicial de venta API puso de manifiesto un patrón muy similar al ya observado en el dataset manual, la superficie es el determinante fundamental del precio, seguida por el efecto negativo de las habitaciones. La planta muestra un posible impacto

positivo que roza la significancia estadística y que merece considerarse en fases posteriores del análisis. El resto de las variables no aportaron evidencia sólida en este primer ajuste, confirmando la necesidad de depurar el modelo para concentrarse en los predictores realmente relevantes.

OLS Regression Results

Dep. Variable:	Precio	R-squared:	0.977
Model:	OLS	Adj. R-squared:	0.975
Method:	Least Squares	F-statistic:	503.6
Date:	Thu, 21 Aug 2025	Prob (F-statistic):	1.42e-64
Time:	18:26:16	Log-Likelihood:	-1271.0
No. Observations:	90	AIC:	2558.
Df Residuals:	82	BIC:	2578.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.093e+05	1.26e+05	-0.866	0.389	-3.61e+05	1.42e+05
Superficie	1.075e+04	282.566	38.046	0.000	1.02e+04	1.13e+04
Habitaciones	-8.578e+04	3.97e+04	-2.158	0.034	-1.65e+05	-6713.260
Baños	-1.196e+04	5.25e+04	-0.228	0.820	-1.16e+05	9.24e+04
Exterior	1.33e+05	1.06e+05	1.255	0.213	-7.78e+04	3.44e+05
Planta	3.614e+04	1.74e+04	2.073	0.041	1457.706	7.08e+04
Aparcamiento	7393.6438	1.04e+05	0.071	0.944	-2e+05	2.14e+05
Ascensor	-3.729e+04	1.1e+05	-0.338	0.736	-2.56e+05	1.82e+05

Omnibus:	20.413	Durbin-Watson:	1.513
Prob(Omnibus):	0.000	Jarque-Bera (JB):	129.146
Skew:	0.171	Prob(JB):	9.04e-29
Kurtosis:	8.858	Cond. No.	1.07e+03

Ilustración 15: Resultados modelo inicial ventas vía API

6.3.2 Optimización mediante eliminación de variables con p-valor alto

Siguiendo la metodología planteada se procedió a depurar el modelo eliminando de manera progresiva las variables con p-valores elevados, es decir, aquellas que no aportaban evidencia estadística suficiente sobre el precio de venta. Quedaron fuera Baños, Exterior, Ascensor y Aparcamiento, dado que todas presentaban p muy superiores a 0,05 en el modelo inicial. Al retirarlas, se recalibró el modelo concentrándose en los predictores más sólidos, Superficie, Habitaciones y Planta.

Antes de presentar los resultados se analizó la correlación entre estos tres predictores, como se observa en la matriz de correlaciones la Superficie mantiene una correlación moderada con el número de Habitaciones ($r \approx 0{,}54$), algo esperable dado que a mayor metraje suelen encontrarse más habitaciones. También se detectó cierta asociación entre Superficie y Planta ($r \approx 0{,}44$), mientras que la correlación entre Habitaciones y Planta fue más débil ($r \approx 0{,}27$). Estos valores confirman que, aunque existe relación entre predictores, no alcanzan niveles

problemáticos de multicolinealidad permitiendo mantenerlos en el modelo sin riesgo excesivo de distorsión en los coeficientes.

Una vez optimizado, el modelo reducido alcanzó un $R^2 = 0,977$ (R^2 ajustado = 0,976), manteniendo prácticamente el mismo nivel de explicación que el modelo inicial (≈ 98%) pero con un conjunto mucho más parsimonioso de variables. Este resultado pone de manifiesto que prácticamente toda la información explicativa se concentra en estos tres predictores y que las variables descartadas añadían ruido y no valor real al ajuste.

Los coeficientes estimados resultaron coherentes con lo esperado, la Superficie se consolidó como el factor más influyente, con un coeficiente de aproximadamente +10.710 € por m² adicional y un $p < 0,001$. Esto confirma de manera rotunda el efecto positivo de la superficie sobre el precio de venta. El número de Habitaciones mostró un coeficiente negativo (≈ – 75.280 € por cada habitación adicional, $p \approx 0,018$), lo que evidencia que, manteniendo constante la superficie, más habitaciones tienden a reducir el valor medio de la vivienda. Se trata de un hallazgo alineado con los análisis previos, la compartimentación en exceso resta atractivo frente a espacios más amplios.

Por último, la Planta arrojó un coeficiente positivo (≈ +37.240 € por nivel adicional) y fue estadísticamente significativo ($p \approx 0,030$). Este resultado confirma que, en este dataset, vivir en pisos más altos se asocia con un precio superior, como previamente explicado, por factores como mayor luminosidad, mejores vistas o menor exposición al ruido. A diferencia del modelo inicial, donde su significancia era solo marginal, en este modelo depurado la Planta superó el umbral del 95% de confianza, lo que justifica plenamente su inclusión.

La mejora observada no se refleja tanto en el incremento del R^2, ya muy elevado desde el inicio, como en la estabilidad de los coeficientes y la reducción de predictores redundantes. Al concentrarse únicamente en Superficie, Habitaciones y Planta, se logra un modelo más claro y robusto, en el que cada variable cumple un papel bien definido, además, el número de observaciones se mantuvo constante, lo cual refuerza aún más la interpretación de que la mejora en estabilidad y significancia se debe únicamente a la depuración de variables irrelevantes.

En resumen, el modelo optimizado de ventas API confirma que la práctica totalidad de la variación en los precios se explica a través de tres elementos fundamentales, el metraje, la distribución interna y la altura en el edificio. Esta combinación permite alcanzar un nivel de ajuste excepcional ($R^2 > 0,97$), constituyéndose como uno de los modelos más consistentes de todo el análisis.

6.3.3 Resultado final con mejor R^2

El resultado final del modelo de ventas vía API, tras la optimización, fue sumamente satisfactorio en términos de ajuste, como se mencionó, $R^2 \approx 0{,}977$ (R^2 ajustado $\approx 0{,}976$), lo cual representa un nivel de explicación extraordinariamente alto y, de hecho, el mejor obtenido entre todos los modelos analizados en este trabajo. Esto indica que el modelo logra predecir con gran precisión los precios de venta usando únicamente un conjunto reducido de variables fáciles de obtener, superficie, número de habitaciones y la planta.

Este desempeño tan destacado invita a reflexionar sobre las razones que lo explican. Varias consideraciones pueden señalarse, cómo la calidad y homogeneidad de los datos, es posible que la muestra de la API corresponda a un segmento del mercado relativamente más homogéneo, donde las relaciones lineales entre características y precio resultan más claras que en el dataset manual, el cual incorporaba viviendas de lujo con rasgos singulares que resultan complejos de modelizar adecuadamente. Una mayor uniformidad de la muestra reduce el ruido y permite detectar con más nitidez el efecto de las variables clave. Otra clave es el tamaño muestral, la base de la API incluyó un número mayor de observaciones (109 propiedades) frente al dataset manual (100). Aunque la diferencia no es enorme, disponer de más datos ayuda a estimar coeficientes más estables y representativos del mercado en general, reduciendo la influencia de valores atípicos.

Continuando con la simplicidad del modelo, el hecho de que el modelo final se haya concentrado en solo tres predictores clave superficie, habitaciones y planta supone una ventaja. Variables adicionales como baños, exterior, ascensor o aparcamiento, que no mostraban significancia, solo introducían ruido, en cambio, al depurarlas, el modelo se volvió más parsimonioso y robusto, maximizando su capacidad predictiva. Por último, el efecto de la depuración, a diferencia de otros casos, el número de observaciones se mantuvo constante en 109 entre el modelo inicial y el reducido, por tanto, la mejora en los resultados no se debe a la exclusión de outliers, sino al hecho de haber eliminado predictores redundantes o poco informativos, esto refuerza la idea de que la ganancia en ajuste es genuina y atribuible a la selección adecuada de variables.

En términos interpretativos la superficie se reafirmó como la variable más influyente, con un coeficiente positivo de alrededor de +10.710 € por cada metro cuadrado adicional. El número de habitaciones ejerció un efecto negativo significativo reduciendo el precio en aproximadamente –75.280 € por cada habitación adicional a igualdad de superficie. Finalmente, la planta tuvo un impacto positivo de unos +37.240 € por nivel, estadísticamente significativo al 95%.

En conclusión, el modelo final con datos de venta vía API logró explicar casi el 98% de la variabilidad de los precios empleando únicamente tres variables principales. Este resultado no solo supera con claridad al modelo equivalente con datos manuales (R^2 ajustado ~0,76), sino también al modelo combinado (R^2 ajustado ~0,88). Ello puede atribuirse a la coherencia interna y representatividad del dataset API. Para validar la solidez de este hallazgo, sería recomendable contrastar este modelo con datos de otros distritos o ciudades, pero dentro del alcance del presente TFG puede afirmarse que se trata del modelo más exitoso en términos de ajuste estadístico y parsimonia.

6.4 Combinación de datos de venta (Extracción manual y API)

Finalmente, se procedió a combinar los datos de venta recopilados manualmente con los datos de venta obtenidos vía API, con el objetivo de explotar un mayor volumen de información y verificar si un modelo global mejora o cambia con respecto a los anteriores. Antes de ajustar el modelo combinado, fue necesario realizar una limpieza para evitar sesgos por registros repetidos.

6.4.1 Eliminación de duplicados

Al unir ambas fuentes de datos (manual y API) se encontró que algunas propiedades aparecían en ambos conjuntos, probablemente porque la recopilación manual y la API incluían ciertos inmuebles en común, anuncios que se habían recogido manualmente de un portal, y que también estaban presentes en la data obtenida por la API del mismo portal. Incluir duplicados sería problemático ya que implicaría dar doble peso a esas observaciones en el ajuste del modelo, por ello, se realizó un filtrado para identificar y eliminar los duplicados antes de ajustar el modelo combinado. La identificación y eliminación de duplicados se realizó de manera automática mediante la función “drop_duplicates” de Python, la cual permite detectar registros repetidos en función de varios atributos comunes (precio, superficie, ubicación, entre otros). De este modo, se garantizó que cada propiedad quedara representada una sola vez en el dataset final, evitando sesgos por repeticiones

Como resultado, el número total de observaciones únicas en el dataset combinado fue N = 175 propiedades (aplicado el 85% de investigación). Este número es algo menor que la suma de los datasets por separado (85 manual + 92 API ≈ 177) precisamente por la superposición de listados que se descartaron. Con los duplicados fuera, se procedió a concatenar los datos y asegurar que las variables consideradas estuvieran presentes para todas las observaciones. Dado que el dataset manual incluía algunas variables (e.g. Terraza, Piscina, Reformado) que la API no proporcionaba, y viceversa, para el modelo combinado se restringió el conjunto de

variables a aquellas comunes o disponibles en ambas fuentes, para evitar vacíos. Es decir, se trabajó esencialmente con las variables estructurales básicas que proporcionaba la API, superficie, habitaciones, baños, planta, ascensor, aparcamiento, exterior. Variables como terraza o piscina, que solo se tenían para las propiedades del subgrupo manual, no podían utilizarse en el combinado completo porque faltarían en muchos casos de la API. Por tanto, el modelo combinado operó en la práctica con el mismo conjunto de predictores que el modelo de venta API y un subconjunto de los del modelo manual original.

6.4.2 Evaluación global del modelo combinado

Se ajustó un modelo de regresión lineal múltiple sobre los 125 registros combinados, utilizando inicialmente las variables comunes mencionadas (Superficie, Habitaciones, Baños, Planta, Ascensor, Aparcamiento, Exterior). El modelo arrojó un $R^2 = 0{,}889$ (R^2 ajustado = 0,882), lo cual refleja un poder explicativo notable teniendo en cuenta la heterogeneidad del conjunto, sin embargo, al analizar los coeficientes y sus p-valores, se constató que únicamente Superficie y Habitaciones eran significativas al nivel del 95%, mientras que el resto de las variables (Baños, Exterior, Ascensor, Aparcamiento e incluso Planta) no alcanzaban significancia estadística.

En cuanto a los coeficientes, la Superficie mostró un valor positivo de aproximadamente +11.270 €/m² y altamente significativo ($p < 0{,}001$), consolidándose como el predictor más relevante y consistente a lo largo de todos los modelos. El número de Habitaciones, por el contrario, presentó un coeficiente negativo de alrededor de –165.800 € por cada habitación adicional y también significativo ($p \approx 0{,}023$), lo que confirma una vez más que, a igualdad de superficie, una mayor compartimentación tiende a reducir el valor percibido de la vivienda. La variable Planta, aunque en otros modelos había mostrado cierta relevancia, en este caso arrojó un coeficiente positivo no significativo ($p > 0{,}1$), por lo que su inclusión no mejoraba el modelo combinado.

El análisis de la matriz de correlaciones ayuda a entender este resultado, variables como Baños (correlación 0,73 con Superficie y 0,68 con Habitaciones) o Aparcamiento (correlación 0,51 con Baños) presentan niveles de asociación relativamente altos con los predictores claves, lo que sugiere redundancia. En la práctica, su efecto ya queda absorbido por la superficie y las habitaciones, de modo que al incluirlas no aportan información nueva y, además, pueden introducir problemas de colinealidad.

En el caso del modelo combinado, desde la primera estimación se observó que únicamente Superficie, Planta y Habitaciones resultaban significativas, mientras que el resto de las variables (Baños, Aparcamiento, Ascensor, Exterior) no mostraban ningún efecto

estadísticamente relevante. Entonces, no se consideró necesario llevar a cabo un proceso de optimización adicional como en otros apartados, el propio modelo inicial ya apuntaba hacia una estructura reducida y parsimoniosa.

Así, el modelo final del conjunto combinado se limitó a esas variables clave, confirmando los hallazgos de los apartados anteriores, los metros cuadrados son el principal determinante del precio, y el número de habitaciones actúa como ajuste corrector con signo negativo, y cuanto más alta la planta mayor el precio. Con solo estos tres predictores se logra explicar alrededor del 89% de la variación total en precios, un resultado muy notable si se tiene en cuenta la heterogeneidad de los datos procedentes tanto de registros manuales como de la API.

En conclusión, la combinación de ambas fuentes de información no alteró las relaciones estructurales ya identificadas, más bien, reforzó la evidencia de que el valor de una propiedad se concentra fundamentalmente en el tamaño y en la manera en que se distribuye ese espacio. Estos hallazgos muestran que incluso en un dataset amplio y diverso, un modelo lineal sencillo permite capturar de manera eficaz los fundamentos del mercado inmobiliario, sin necesidad de incorporar predictores secundarios que no aportan significancia estadística.

6.5 K-means

En una zona urbana heterogénea como Nuevos Ministerios - Ríos Rosas, el mercado inmobiliario ofrece viviendas de muy diversas características (tamaños, precios, antigüedad, etc.), para comprender mejor estos patrones de oferta y detectar tipologías de vivienda predominantes, resulta útil emplear técnicas de clustering como K-means. Este algoritmo agrupa automáticamente las viviendas en segmentos homogéneos según sus características, revelando perfiles o nichos de mercado difíciles de identificar a simple vista. En otras palabras, K-means permite clasificar las viviendas en grupos de propiedades similares, por ejemplo, pisos económicos frente a viviendas de alta gama de forma objetiva. Estudios previos han utilizado K-means para segmentar mercados residenciales de manera similar, clasificando las viviendas en categorías como económicas, medias o lujosas en base a atributos comunes. Al aplicar este enfoque en Nuevos Ministerios - Ríos Rosas se busca descubrir patrones significativos que orienten estrategias comerciales y aporten una comprensión más profunda del mercado local.

Tras aplicar K-means por separado a los datos de viviendas en venta y en alquiler de la zona se obtuvieron varios clusters o grupos de viviendas con perfiles bien diferenciados. A continuación, se describen las características principales de los grupos identificados en cada

caso, así como su posible interpretación dentro de la oferta inmobiliaria de Nuevos Ministerios - Ríos Rosas.

6.5.1 Perfiles de viviendas en venta

En la segmentación de ofertas de venta el algoritmo identificó esencialmente cinco perfiles predominantes de vivienda en la zona. Cada grupo comparte rasgos comunes en cuanto a precio, tamaño y prestaciones de las viviendas.

```
Medianas numéricas por perfil:
              Precio  Superficie  Habitaciones  Baños  Planta
perfil
0        1107500.0       116.0           3.0    2.0     2.0
1       15000000.0      1390.5           5.5    6.5    10.0

Proporciones (binarias) por perfil:
         Aparcamiento  Exterior
perfil
0               0.226     0.821
1               1.000     1.000
```

Ilustración 16: Perfiles K-Means venta

Viviendas de alto precio y gran tamaño (lujosas).

Uno de los clusters agrupa pisos amplios y caros, generalmente exteriores y con múltiples comodidades, en este grupo se encuentran las viviendas de mayor superficie (por ejemplo, varias habitaciones y 100+ m²) y calidades altas, típicamente ubicadas en las mejores ubicaciones del barrio. Muchas de estas propiedades además incluyen plaza de garaje, un equipamiento valioso y escaso en la zona, lo que eleva su precio. Este perfil corresponde a pisos de lujo para familias o clientes de alto poder adquisitivo que buscan espacio y confort completos.

Viviendas de alto precio y tamaño reducido.

Curiosamente, otro grupo identifica pisos pequeños pero caros. Son apartamentos o estudios de superficie limitada (por ejemplo, 1 dormitorio o menos de 60 m² de superficie) cuyo precio por metro cuadrado es muy alto. Suelen ser inmuebles exteriores y frecuentemente nuevos o reformados de lujo, a veces con extras como garaje o trastero pese a su tamaño compacto. Este perfil sugiere la existencia de viviendas “premium” de menor dimensión, dirigidas quizá a inversores, parejas jóvenes o profesionales que valoran la ubicación y calidades modernas por encima del espacio, es decir, pisos pequeños en ubicaciones exclusivas del barrio que alcanzan precios elevados por sus acabados o por estar en edificios representativos.

Viviendas económicas de tamaño pequeño.

En contraste con los anteriores, también aparece un cluster de pisos asequibles y compactos, son las viviendas de menor precio relativo en la zona típicamente de poca superficie (estudios o apartamentos de 1-2 dormitorios) y características modestas. Dentro de este grupo exceden los inmuebles sencillos, a menudo en edificios antiguos o con menos atractivos. Muchas de estas unidades son interiores o en plantas bajas, lo cual reduce su valor, aunque algunas siguen siendo exteriores, pero con peor calidad. Este perfil representa la oferta de entrada en Nuevos Ministerios - Ríos Rosas, viviendas para compradores con presupuesto limitado, inversores que buscan piso pequeño para alquilar, o como primera vivienda asequible en la zona.

Viviendas de alto precio, gran tamaño y con aparcamiento.

Una segmentación más fina dentro de las viviendas caras de gran tamaño distingue aquellas que sí disponen de garaje. Este subgrupo de pisos exclusivos combina amplitud, lujo y aparcamiento, reuniendo propiedades de alta gama ideales para familias acomodadas que buscan todas las comodidades en pleno centro. Aunque conceptualmente se solapa con el primer perfil descrito de "gran tamaño y caro", se separa específicamente por el factor aparcamiento, lo que indica que en este mercado tener garaje propio marca una diferencia clara dentro de la oferta de lujo.

Viviendas de alto precio, gran tamaño y sin aparcamiento.

Otro cluster agrupa los pisos grandes y caros que no cuentan con garaje, estos inmuebles mantienen características de lujo (espaciosos, exteriores y buena ubicación), pero carecen de ese extra. Su existencia como grupo diferenciado sugiere que incluso entre las viviendas de precio elevado hay segmentación según equipamientos, en Nuevos Ministerios - Ríos Rosas disponer o no de garaje puede posicionar a una vivienda en submercados distintos, reflejando la valoración extra que el mercado da a este servicio en ventas de alto nivel.

En resumen, los clusters de venta dibujan por un lado un segmento de lujo (pisos amplios y caros, subdividido según tengan o no garaje), por otro lado, un segmento premium compacto (pisos pequeños pero muy costosos), y finalmente un segmento económico de pisos pequeños y baratos. Prácticamente todas las agrupaciones resaltan la diferencia entre propiedades caro vs barato y grande vs pequeño, incluyendo también el carácter exterior (con luz natural) como atributo valioso mencionado en la mayoría de los perfiles. Esto concuerda con la intuición de que en esta zona hay un contraste marcado entre viviendas exclusivas y viviendas más modestas conviviendo en el mismo mercado.

6.5.2 Perfiles de viviendas en alquiler

Para las ofertas de alquiler residencial en la zona, el análisis K-means también identificó aproximadamente cinco grupos principales con similitudes y matices distintos respecto a las tipologías de venta.

```
Medianas numéricas por perfil:
              Precio  Superficie  Habitaciones  Baños  Planta
perfil
0          1107500.0       116.0           3.0    2.0     2.0
1         15000000.0      1390.5           5.5    6.5    10.0

Proporciones (binarias) por perfil:
         Aparcamiento  Exterior
perfil
0               0.226     0.821
1               1.000     1.000
```

Ilustración 17: Perfiles K-means alquiler

Pisos pequeños de alquiler asequibles.

Uno de los clusters corresponde a las unidades de alquiler más baratas y de superficie reducida, son estudios y apartamentos de pocos metros cuadrados con renta mensual baja para los estándares del barrio. Suelen ser viviendas simples, a menudo interiores o con menos luz, ideales para estudiantes o parejas jóvenes que buscan vivir en la zona por un precio asequible. Este perfil es equivalente al de viviendas económicas en venta, reflejando que también en alquiler existe un segmento de oferta básica para presupuesto limitado.

Pisos grandes de alquiler de precio elevado.

En el otro extremo se encuentra un cluster de viviendas amplias con rentas altas, aquí se agrupan los pisos de mayor superficie en alquiler (por ejemplo, 3 o más dormitorios), ubicados en buenas fincas y con rentas mensuales muy elevadas. Suelen ser propiedades exteriores, luminosas y de calidad, pensadas para familias, directivos o internacionales dispuestos a pagar más a cambio de mayor espacio y mejor localización. Este segmento equivale al de lujo en venta, identifica la oferta premium en alquiler, marcada por el gran tamaño y precio. Cabe señalar que, a diferencia de la venta, en este grupo la disponibilidad de garaje no aparece tan destacada, muchas viviendas de alquiler grandes pueden no incluir aparcamiento y aun así pertenecer al segmento caro por su tamaño y ubicación.

Pisos pequeños de alquiler con rentas altas.

Similar al caso de venta, también en alquiler surge un perfil de apartamentos de escasa superficie, pero alquiler elevado. Son estudios o pisos de 1 dormitorio, muy modernos o bien ubicados, cuyos propietarios piden rentas relativamente altas por metro cuadrado, este perfil

indica la presencia de una demanda para viviendas pequeñas pero exclusivas en alquiler. Estas unidades, aunque sean pequeñas logran rentas comparables a viviendas mayores debido a sus calidades o localización privilegiada.

Pisos grandes en alquiler a precio moderado con aparcamiento.

Un cluster particular del mercado de alquiler reúne viviendas de gran tamaño con alquiler sorprendentemente más bajo con relación a su superficie, y muchos de ellos incluyen garaje. Este hallazgo sugiere que existen pisos amplios (quizá en edificios menos nuevos o en ubicaciones más excéntricas dentro del barrio) cuyos propietarios ofrecen rentas competitivas. La inclusión de un aparcamiento en estos casos podría indicar edificios más modernos, pero no en las zonas más cotizadas de Nuevos Ministerios - Ríos Rosas, o bien propiedades donde el precio por m² es bajo por ser antiguas, necesitadas de reforma, etc. Este perfil no tuvo equivalente claro en la segmentación de venta, ya que en ventas cualquier piso grande tiende a ser costoso, pero en alquiler se observan algunos chollos de mayor tamaño que destacan como un grupo separado. Posiblemente corresponden a propietarios que priorizan asegurar un inquilino estable a un precio razonable, o viviendas familiares más antiguas que no alcanzan rentas de lujo pese a su tamaño.

6.5.3 Diferencias entre perfiles

Al comparar los clusters de venta y alquiler, se aprecian similitudes en la existencia de perfiles análogos, pero también diferencias significativas dictadas por la naturaleza de cada mercado. En venta la presencia o ausencia de garaje resultó ser un factor muy diferenciador entre viviendas de alta gama, creando sub-perfiles específicos. En cambio, en alquiler el factor garaje tiene menos peso segmentando el mercado, solo un cluster distintivo (pisos grandes moderados con aparcamiento) lo refleja, mientras que en los alquileres más caros el aparcamiento no separa grupos con tanta claridad. Esto sugiere que los compradores valoran enormemente una plaza de aparcamiento, impactando fuertemente el precio de venta, mientras que los arrendatarios pueden estar más dispuestos a prescindir de ella o conseguirla por otras vías, priorizando más la superficie o la ubicación en el precio que pagan mensualmente.

Por otro lado, se observa en alquiler un fenómeno de pisos grandes baratos relativo a su tamaño, algo que no aparece en venta. Esto implica que algunas viviendas de gran metraje en la zona se logran alquilar a precios por debajo de lo esperado, formando un segmento propio. En venta cualquier vivienda de gran tamaño en el barrio prácticamente conlleva un precio elevado, de modo que no se formó un cluster de grandes baratos, esto refleja dinámicas distintas. El mercado de compraventa tiende a valorar el metro cuadrado de forma más

consistente y los pisos grandes siempre cotizan alto, mientras que en alquiler puede haber propietarios flexibilizando rentas por diversos motivos, creando oportunidades atípicas en ese segmento.

Ambos mercados comparten un segmento económico de viviendas pequeñas, pero el peso y composición de este grupo podría diferir, las viviendas en venta baratas y pequeñas podrían incluir más interiores o unidades a reformar, como inversión, mientras que en alquiler las viviendas pequeñas baratas incluyen estudios sencillos a precios asequibles para inquilinos locales. En general, el perfil de vivienda básica existe tanto en venta como alquiler, pero responde a perfiles distintos, compradores de presupuesto limitado e inquilinos buscando rentas bajas, respectivamente.

En cuanto a las viviendas pequeñas de alto precio, tanto en venta como en alquiler se detectan clusters que representan un nicho de lujo (pisos de tamaño reducido pero muy cotizados), sin embargo, su presencia puede ser más notoria en venta, donde pagar mucho por pocos metros puede deberse a inversión en zonas más demandadas que en alquiler, donde existe un techo en lo que un inquilino pagaría por un piso muy pequeño. Aun así, la identificación de este perfil en ambos mercados confirma una tendencia incluso en superficies pequeñas, si la ubicación y calidad son excelentes se genera una oferta de alto nivel que encuentra su demanda.

En resumen, la segmentación revela que, aunque venta y alquiler comparten tipologías generales de vivienda, las prioridades de compradores y de arrendatarios introducen variaciones en cómo se agrupan esas viviendas. Los compradores tienden a segmentar más por valor a largo plazo y equipamientos, mientras que los inquilinos optan más por coste mensual y flexibilidad, lo que puede dar lugar a oportunidades como grandes pisos en alquiler relativamente económicos, algo inexistente en venta. Estas diferencias son valiosas para entender las dinámicas específicas de cada submercado dentro de Nuevos Ministerios - Ríos Rosas.

6.5.4 Conclusiones

El uso de K-means para identificar perfiles de vivienda aporta un gran valor práctico en la comprensión del mercado residencial local, para empezar, permite sintetizar grandes volúmenes de datos inmobiliarios (precios, tamaños y otras características de las viviendas) en unos pocos segmentos interpretables cada uno con un perfil definido. Esto facilita reconocer, por ejemplo, qué caracteriza a una vivienda típica de lujo en la zona frente a una vivienda estándar o económica. Estudios previos ya señalaban que los grupos resultantes

presentan diferencias significativas en términos de precio, superficie y ubicación, lo cual posibilita separar el mercado de manera eficiente. Dicho de otro modo, modelizar con clusters convierte datos dispersos en información accionable, se identifican nichos de mercado concretos y sus atributos clave, lo que ayuda a enfocar estrategias.

Para los agentes inmobiliarios estos hallazgos permiten personalizar la comercialización, conociendo los perfiles, un agente puede ajustar el mensaje de venta o alquiler al segmento objetivo, por ejemplo, destacar el lujo y exclusividad al promocionar un piso del cluster alta exclusividad o resaltar la oportunidad de precio al ofrecer una vivienda del cluster económico pequeño. También ayuda en la tasación y negociación, saber en qué segmento cae una propiedad orienta sobre su valor relativo en el mercado, si perteneciese al grupo top, al medio o al básico, y qué tipo de comprador o inquilino buscará, así, las estrategias de marketing y precios se afinan según el segmento identificado.

Para los promotores y desarrolladores inmobiliarios, el análisis de clusters ofrece una visión para la planificación de proyectos, entender qué tipologías abundan en Nuevos Ministerios - Ríos Rosas y cuáles quizás están subofertadas puede guiar decisiones sobre futuros desarrollos. Por ejemplo, si se detecta que casi todas las viviendas nuevas son de lujo pequeño y falta producto de tamaño medio para familias a precio razonable, un promotor podría orientar su próxima promoción a llenar ese vacío de mercado, y del mismo modo para pisos en alquiler. En resumen, la segmentación ayuda a identificar oportunidades, ya sea explotar un nicho existente o diferenciar la oferta para dirigirse a un segmento específico.

Finalmente, para cualquier interesado en el mercado residencial local, desde analistas urbanos hasta compradores informados, esta modelización ofrece una mejor comprensión de la estructura de precios y calidades en el barrio. Se hace evidente qué factores (espacio, ubicación, extras como garaje) explican la diferencia de precios y cómo se agrupan las viviendas en submercados relativamente homogéneos. Esto contribuye a una transparencia mayor, por ejemplo, un comprador puede entender si la vivienda que busca pertenece al segmento económico o premium y ajustar sus expectativas. En conclusión, la técnica de K-means aplicada al mercado inmobiliario de Nuevos Ministerios - Ríos Rosas enriquece el análisis estratégico, proporcionando un mapa claro de los perfiles de vivienda existentes, lo que redunda en decisiones más informadas tanto para la oferta como para la demanda

7. Conclusiones y trabajos futuros

7.1 Conclusión

El presente Trabajo de Fin de Grado ha analizado el precio de la vivienda en la zona Nuevos Ministerios–Ríos Rosas (Madrid) mediante técnicas de análisis exploratorio, regresión lineal y clustering K-means, aplicadas sobre datos de alquiler y venta obtenidos desde la web de Idealista a través de una API inmobiliaria. A la luz de los resultados obtenidos se pueden extraer varias conclusiones relevantes, para empezar, se confirma el papel determinante de la superficie de la vivienda, del número de habitaciones y de la planta en la explicación del precio. Estas variables emergen como los factores con mayor influencia superando a otros atributos en poder explicativo, en otras palabras, las viviendas más amplias y con más estancias tienden sistemáticamente a precios mayores, lo cual concuerda con la intuición y la evidencia común del mercado inmobiliario. Este hallazgo refuerza la idea de que en la zona estudiada el tamaño, la capacidad habitacional y la altura de la vivienda son los principales motores de su valor.

En segundo lugar, los modelos implementados han demostrado una notable capacidad para explicar la variabilidad del precio, aunque con diferencias según la técnica empleada. El modelo de regresión lineal múltiple, que cuantifica la relación entre las características de las viviendas y el precio, logró un ajuste significativo, evidenciando que los atributos considerados explican una proporción amplia de la variación observada en los precios. Esto indica que los datos recopilados contienen la información suficiente para estimar el precio con bastante precisión mediante una ecuación lineal. Por su parte, el algoritmo de agrupamiento K-means, aunque no es un modelo predictivo, permitió identificar patrones naturales o segmentos en los datos, agrupando las viviendas en clústeres relativamente homogéneos. Estos clústeres revelan diferencias significativas en los perfiles de las viviendas (por ejemplo, grupos de pisos pequeños económicos frente a pisos de lujo de gran tamaño), lo cual aporta una perspectiva complementaria a la de la regresión. En conjunto, la aplicación combinada de técnicas complementarias ha enriquecido el análisis, la regresión proporciona una medida cuantitativa global del efecto de cada variable sobre el precio, mientras que el clustering captura matices segmentados y posibles relaciones no lineales entre atributos que el modelo lineal al promediar los efectos podría haber pasado por alto.

Un ejemplo ilustrativo de las diferencias de enfoque entre la regresión lineal y K-means es el papel de la plaza de aparcamiento, el modelo de regresión lineal no detectó un impacto estadísticamente significativo de la disponibilidad de garaje en el precio de la vivienda, lo que sugiere que dentro de la muestra analizada este factor queda eclipsado por otros más

determinantes o bien su efecto medio no resulta lo bastante grande para destacar en la regresión. Sin embargo, el análisis de clustering sí puso de manifiesto que la presencia o ausencia de aparcamiento marca una distinción entre ciertos grupos de viviendas. En los resultados de K-means se observó que las propiedades con plaza de garaje tendieron a agruparse en clústeres diferenciados de aquellas que no disponen de ella, esto indica que aunque el efecto del garaje no se refleje de forma aislada en el modelo lineal global, en la segmentación del mercado este atributo sí juega un papel importante, las viviendas con aparcamiento conforman un subconjunto con características de precio y prestaciones algo diferentes a las del subconjunto de viviendas sin aparcamiento. Dicho hallazgo pone de relieve cómo diferentes enfoques analíticos pueden ofrecer conclusiones complementarias, la regresión resume la influencia promedio de cada variable en toda la muestra, mientras que la segmentación con K-means descubre heterogeneidades internas en el mercado, destacando factores, como el garaje, que pueden ser relevantes para determinados segmentos de viviendas, aunque su influencia global parezca tenue.

En conclusión, las técnicas aplicadas han permitido cumplir con el objetivo de caracterizar y modelizar el precio de la vivienda en Nuevos Ministerios–Ríos Rosas, identificando claramente los determinantes principales y cuantificando su efecto. Al mismo tiempo, la comparación entre modelos ha evidenciado las fortalezas y limitaciones de cada enfoque, la regresión lineal ofrece simplicidad e interpretabilidad en la estimación del precio, mientras que el clustering K-means aporta una visión estratégica sobre cómo se estructura el mercado en submercados o nichos. Estas conclusiones proporcionan una comprensión más completa del fenómeno estudiado y sirven de base para orientar decisiones en contextos reales, como estrategias de marketing inmobiliario, en el ámbito local analizado.

7.2 Trabajos futuros

Partiendo de los hallazgos anteriores se identifican varias direcciones en las que podría extenderse o profundizar este trabajo en el futuro, una propuesta especialmente innovadora sería la simulación de anuncios inmobiliarios con características ficticias para validar las estimaciones de precio del modelo en un entorno real. Con los coeficientes obtenidos de la regresión lineal, se podría tomar un conjunto de características hipotéticas (por ejemplo, un piso de X metros cuadrados, con Y habitaciones, con o sin garaje, etc) y calcular el precio estimado que el modelo asignaría a un inmueble así. A continuación, se podría publicar ese anuncio simulado en portales inmobiliarios reales como si se tratase de una vivienda en alquiler o venta, observando cómo responde el mercado ante dicho precio. El objetivo de este

experimento sería medir de forma empírica la validez y competitividad de los precios sugeridos por el modelo, si el modelo predice correctamente el valor de mercado, cabría esperar que el anuncio generase interés en un plazo razonable, por el contrario, si el precio estimado está desajustado (demasiado alto, por ejemplo), es probable que el anuncio reciba poca atención o tarde mucho en encontrar interesado.

Para evaluar el éxito o la respuesta de estos anuncios simulados se pueden definir métricas concretas obtenibles a través de los propios portales inmobiliarios. Por ejemplo, el número de visitas que registra el anuncio en la web sería un primer indicador del interés que genera. Asimismo, podría medirse el número de contactos o consultas realizados por potenciales clientes, personas que solicitan información adicional o una visita al inmueble. Finalmente, el indicador más claro sería el tiempo hasta la reserva o cierre, es decir, cuánto tiempo transcurre antes de que el anuncio sea retirado por haberse reservado o vendido la vivienda, en el caso de un anuncio simulado podría medirse el tiempo hasta recibir una oferta firme o hasta que un volumen significativo de interesados se apunte. Un anuncio correctamente valorado, precio acorde al mercado, probablemente recibiría numerosas visitas y contactos en pocos días y podría marcarse como reservado en un plazo corto, mientras que un anuncio con un precio excesivo permanecería más tiempo sin apenas consultas. Esta aproximación experimental, además de original, brindaría una forma de validación externa del modelo de regresión, ya que se comprobaría en la práctica si las predicciones de precio se traducen en transacciones ágiles en el mercado real.

Adicionalmente, existen otras posibles extensiones de este trabajo que podrían mejorar o ampliar los resultados obtenidos, una de ellas es incluir variables explicativas adicionales que no estaban disponibles en el conjunto de datos original pero que podrían influir en el precio de forma importante. Por ejemplo, características como el estado de conservación o reforma de la vivienda (si ha sido renovada recientemente o si necesita mejoras), la antigüedad del edificio, si tiene terraza, la existencia de zonas comunes (jardines, trastero, portero) o incluso indicadores del entorno (nivel de ruido, proximidad a transporte público, oferta comercial cercana) son factores que la literatura inmobiliaria identifica como relevantes en la formación de precios. Incluir este tipo de variables en futuros análisis permitiría refinar el modelo de pricing y explicar parte de la variabilidad del precio que en el presente estudio queda como residual u oculta por falta de datos. De igual modo, podría contemplarse ampliar el alcance geográfico del estudio, analizar otras zonas de Madrid, u otras ciudades, aplicando la misma metodología serviría para contrastar si los patrones identificados en Nuevos Ministerios–Ríos Rosas son generales o peculiares de este barrio. Cada zona urbana tiene sus propias dinámicas de oferta y demanda, por ejemplo, es posible que en áreas periféricas la disponibilidad de aparcamiento tenga un efecto mucho más marcado en el precio, o que en zonas de mayor

tamaño geográfico la localización exacta dentro del barrio introduzca variaciones significativas. Comparar resultados entre zonas proporcionaría una validación externa geográfica y podría descubrir diferencias locales en los determinantes del precio, enriqueciendo así las conclusiones del trabajo.

Finalmente, otra línea futura de gran interés sería la exploración de modelos de predicción más complejos que complementen o mejoren la aproximación de la regresión lineal. Si bien la regresión lineal ofreció en este proyecto un modelo interpretable y de buen rendimiento global, existen algoritmos de machine learning más avanzados que podrían capturar relaciones no lineales o interacciones sutiles entre las variables. Por ejemplo, podrían probarse árboles de decisión y métodos basados en árboles (random forests, XGBoost u otros modelos de boosting), así como técnicas de redes neuronales o support vector machines. Estos enfoques suelen manejar mejor la complejidad cuando el comportamiento del precio depende de combinaciones específicas de factores, por ejemplo, que el efecto de tener terraza dependa de la planta del piso y de la zona, algo difícil de modelar linealmente. La aplicación de tales modelos, acompañada de una rigurosa validación cruzada y un análisis de su interpretabilidad, para no perder la capacidad de explicar el porqué de cada estimación, podría aumentar la precisión predictiva y proporcionar nuevas perspectivas. En particular, métodos como los basados en árboles pueden destacar qué variables o umbrales dividen más claramente el espacio de decisiones, por ejemplo, puede descubrir automáticamente que por encima de cierta superficie el efecto precio se dispara, o que existen segmentos de mercado bien definidos por rangos de características. Cualquier mejora en la capacidad predictiva o explicativa redunda en un modelo de pricing más útil para agentes reales, ya sea para estimar valor de mercado con menos error o para entender mejor los factores críticos que diferencian unas viviendas de otras.

Por último, cabe señalar que, más que recomendar exclusivamente el uso de una API para la recogida de datos, resulta fundamental plantear un estudio más longevo en el tiempo que permita captar con mayor amplitud la dinámica del mercado inmobiliario. En el caso analizado, el número de observaciones resulta limitado por la propia dimensión de la zona de estudio, lo que refuerza la importancia de prolongar la toma de datos para obtener un análisis más robusto. Además, se ha comprobado que en un horizonte corto la recolección manual de datos puede ser incluso más ágil que la automatización mediante programación, especialmente si no se cuenta con experiencia técnica previa. No obstante, a medida que se extiende el periodo de estudio, la implementación de herramientas de programación y automatización de consultas sí se convierte en una estrategia más eficiente, al permitir recopilar y actualizar información de forma sistemática y con menor esfuerzo operativo.

En conclusión, este proyecto ha logrado identificar los factores clave que explican el precio de la vivienda en la zona estudiada y ha demostrado cómo combinar técnicas tradicionales y modernas de análisis puede aportar un conocimiento más profundo del mercado inmobiliario local. Las conclusiones obtenidas, en especial la primacía de la superficie como determinante del precio, y la evidencia de patrones diferenciados como el impacto del garaje según la segmentación, contribuyen tanto a la literatura existente como a la práctica profesional en tasación y gestión inmobiliaria. Por otro lado, las líneas futuras de trabajo propuestas abren oportunidades para ampliar y aplicar estos hallazgos, desde validar el modelo en entornos reales mediante experimentos con anuncios simulados, hasta enriquecer el análisis con nuevos datos y técnicas más sofisticadas. De este modo, el presente TFG sienta unas bases sólidas sobre las que se puede seguir investigando y mejorando la comprensión del precio de la vivienda, ya sea en Nuevos Ministerios–Ríos Rosas o en contextos más generales, contribuyendo al objetivo último de optimizar el pricing residencial con rigor técnico y visión integral.

8. Bibliografía

Ambrosio Flores, L., Iglesias Martínez, L., Marín Ferrer, C., Pascual Gallego, V., & Serrano Bermejo, A. (2007). A multiple indicators and multiple causes model for valuation: the case of agricultural land. Spanish Journal of Agricultural Research, 5(2), 119-129.

Arellano, M. (2003). Panel Data Econometrics. Oxford: Oxford University Press.

Baltagi, B. H. (2005). Econometric Analysis of Panel Data (3rd ed.). Chichester: John Wiley & Sons.

Banco de España (2003). Análisis del precio de la vivienda en España. Documento de Trabajo Nº 0307, Madrid: Banco de España. https://www.bde.es/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosTrabajo/03/Fic/dt0307.pdf

Baryla, E. A., & Zumpano, L. V. (1995). Residential real estate prices and school quality: Using GIS and hedonic models. Journal of Real Estate Research, 10(1), 1–8.

Brown, G. R., & Matysiak, G. A. (2000). Real estate investment: A capital market approach. Harlow: Pearson Education Limited.

Brueggeman, W. B., & Fisher, J. D. (2011). Real estate finance and investments (14th ed.). New York: McGraw-Hill/Irwin.

Caridad y Ocerín, J. M., & Brañas, P. (1996). Demanda de características de la vivienda en Córdoba: un modelo de precios hedónico. Revista de Estudios Regionales, (46), 139-153.

Caridad y Ocerín, J. M., & Ceular, N. (2001). Un análisis del mercado de la vivienda a través de sistemas de redes neuronales. Revista de Estudios de Economía Aplicada, 18, 67-81.

De los Bueis Villa, S. (2018). Análisis del precio de la vivienda en propiedad y en alquiler en la ciudad de Valladolid. Trabajo de Fin de Grado, Universidad de Valladolid. https://uvadoc.uva.es/handle/10324/34071

Glaeser, E. L., & Gyourko, J. (2005). Urban decline and durable housing. Journal of Political Economy, 113(2), 345–375.

Houlié, N. (2025). The Impact of Economic Policies on Housing Prices: Approximations and Predictions in the UK, the US, France, and Switzerland from the 1980s to Today. Risks, 13(5), 81. https://doi.org/10.3390/risks13050081

Hsiao, C. (2014). Analysis of Panel Data (3rd ed.). Cambridge: Cambridge University Press.

International Valuation Standards Council (IVSC) (2005). International Valuation Standards. London: IVSC.

Jiménez López, F. (2018). Factores determinantes de la diferencia de precios en el mercado inmobiliario. Caso aplicado a los municipios de Cataluña. Trabajo de Fin de Grado, Universitat de Barcelona.

https://diposit.ub.edu/dspace/bitstream/2445/126381/1/TFG-ECO-Jimenez-Francesc-Jun18.pdf

Ling, D. C., & Archer, W. R. (2018). Real estate principles: A value approach (6th ed.). New York: McGraw-Hill Education.

Muñoz, M., & Rubiera, F. (2021). Bayesian networks for housing market analysis: Evidence from Madrid, Barcelona and Valencia. Journal of Housing and the Built Environment, 36(3), 795–818.

Montero Alonso, C. (2020). Las variables físicas del precio del alquiler: mecanismos de vinculación al soporte físico. Trabajo Fin de Grado, Universidad Politécnica de Madrid.

Poterba, J. M. (1984). Tax Subsidies to Owner-Occupied Housing: An Asset-Market Approach. Quarterly Journal of Economics, 99(4), 729-752.

Pudas, L. (2023). Data-driven real estate valuation: A machine learning approach. Bachelor's Thesis, LUT University.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. Journal of Political Economy, 82(1), 34-55.

Shiller, R. J., & Weiss, A. N. (1999). Evaluating Real Estate Valuation Systems. Journal of Real Estate Finance and Economics, 18(2), 147-161.

Sirmans, G. S., Macpherson, D. A., & Zietz, E. N. (2005). The Composition of Hedonic Pricing Models. Journal of Real Estate Literature, 13(1), 3-43.

Thomsett, M. C., & Kahr, J. (2005). Real Estate Market Valuation and Analysis. New York: John Wiley & Sons.

Velasco, J. (2021). Qué elementos explican el precio de la vivienda en alquiler: Estudio del caso de Oviedo. Trabajo de Fin de Grado, Universidad de Oviedo.

Wooldridge, J. M. (2010). Econometric Analysis of Cross Section and Panel Data (2nd ed.). Cambridge, MA: MIT Press.

Yiu, C. Y., & Xu, S. (2012). A Generalized Additive Model for Mass Appraisal of Residential Units. Journal of Property Research, 29(4), 310-328.

Catastro. (s. f.). *Sede electrónica del Catastro*. Ministerio de Hacienda y Función Pública. https://www.sedecatastro.gob.es

Idealista. (s. f.-a). *API de Idealista para desarrolladores*. Idealista. https://developers.idealista.com/access-request

Idealista. (s. f.-b). *API de comparables y métricas*. Idealista Data. https://www.idealista.com/data/asesoramiento-inmobiliario-tecnologico/api-comparables-y-metricas

Idealista. (s. f.-c). *Herramienta de estimación de precio de vivienda*. Idealista. https://www.idealista.com

Instituto Nacional de Estadística (INE). (s. f.). *Estadísticas del mercado inmobiliario*. INE. https://www.ine.es

Lobstr.io. (s. f.). *Lobstr.io – Web scraping platform*. Lobstr.io. https://www.lobstr.io

Octoparse. (s. f.). *Octoparse: Web scraping tool*. Octoparse. https://www.octoparse.com

UrbanData Analytics. (s. f.). *Soluciones de big data inmobiliario*. UrbanData Analytics. https://www.urbandataanalytics.com

Lee, S. (2025, junio 12). *Holdout Method: A Simple yet Effective Technique for Evaluating Machine Learning Models*. *Number Analytics*. Recuperado de

https://www.numberanalytics.com/blog/holdout-method-evaluating-machine-learning-models

Huang, Z., & Lai, G. (2023). A House Price Prediction Model Based on K-means Clustering and Random Forest in Guangzhou. Frontiers in Business Economics and Management, 10(2), 377-381. (Clasificación de viviendas en categorías económicas, confortables y de alta gama).

Villada-Molina, D., et al. (2022). A House Price Modeling Based on Clustering and Kriging: The Medellín Case. The Review of Regional Studies, 52(3), 323-348. (Agrupación de viviendas en submercados locales mediante K-means).

Análisis de Mercado Inmobiliario (2023). Segmentación de viviendas con K-Means. [Informe R-Studio Pubs]. (Identificación de segmentos homogéneos por precio, superficie y ubicación; utilidad en estrategias de marketing).

9. Anexo

```
BASE_URL   = "https://api.idealista.com/3.5/es/search"
COUNTRY    = "es"
LANG       = "es"
OPERATION  = "rent"        # 'sale' o 'rent'
PROP_TYPE  = "homes"       # 'homes', 'offices', etc.
MAX_ITEMS  = 50            # la API máx 50 por página
SORT       = "asc"         # 'asc' o 'desc'

CENTER     = "40.4424,-3.6974" # coordenadas del centro del distrito
DISTANCE_M = "500" #distancia en m al centro (radio)

MAX_PAGES_TO_FETCH = 5  # evita llegar a 100 por seguridad

def build_params(num_page: int) -> dict:
    return {
        "operation": OPERATION,
        "propertyType": PROP_TYPE,
        "language": "es",
        "country": "es",
        "center": CENTER,
        "distance": DISTANCE_M,
        "maxItems": 50,
        "numPage": num_page,
        "order": "publicationDate",
        "sort": SORT,
    }

def search_page(bearer_token: str, url: str, numPage: int) -> Dict[str, Any]:
    """
    Lanza la búsqueda (POST) para una página concreta.
    """
    headers = {"Authorization": f"Bearer {bearer_token}"}
    resp = requests.post(url, headers=headers, data=build_params(numPage))
```

```
    if resp.status_code != 200:
        raise RuntimeError(f"Error search ({resp.status_code}): {resp.text}")
    return resp.json()


def fetch_listings() -> List[Dict[str, Any]]:
    """
        Descarga los anuncios paginando hasta agotar páginas o alcanzar
        MAX_PAGES_TO_FETCH.
    """
    first_page = search_page(TOKEN, BASE_URL, 1)
    total_pages_api = int(first_page.get("totalPages", 1))
    total_pages = min(total_pages_api, MAX_PAGES_TO_FETCH)
    print(total_pages)


    all_elems = []
    all_elems.extend(first_page.get("elementList", []))


    for p in range(2, total_pages + 1):
        page_data = search_page(TOKEN, BASE_URL, p)
        elems = page_data.get("elementList", [])
        all_elems.extend(elems)
        time.sleep(1)


    return all_elems


def to_csv(elements: List[Dict[str, Any]], csv_path: str) -> None:
    """
    Normaliza el JSON a tabla y guarda a CSV.
    """
    if not elements:
        print("No se obtuvieron anuncios.")
        return
    df = pd.json_normalize(elements)
    df.to_csv(csv_path, index=False, encoding="utf-8")
```

```
print(f"Guardado CSV con {len(df)} filas en: {csv_path}")


if __name__ == "__main__":
    elements = fetch_listings()
    to_csv(elements, "idealista_listings7.csv").csv")
```

Código 1: Extracción de datos de la API

```
import numpy as np
import pandas as pd
import statsmodels.api as sm
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
from statsmodels.stats.outliers_influence import variance_inflation_factor
import matplotlib.pyplot as plt
import seaborn as sns


ruta_excel = "prueba.xlsx"


# Si columna de precio tiene otro nombre, cámbiala aquí
df = pd.read_excel(ruta_excel)
df


feature_columns = [
    "Precio (€)", "Superficie (m2)", "Habitaciones", "Baños",
    "Reformado", "Exterior", "Ascensor",
    "Terraza", "Aire acondicionado",
    "Planta", "Aparcamiento"
    "Amueblado", "Piscina"
]


# Nos quedamos con las que existan en el archivo
df = df[feature_columns]
```

```
# Usado solo para datos de API
# df["Aparcamiento"] = df["Aparcamiento"].fillna(False)

df = df.drop_duplicates()
df = df.dropna()

bool_cols = df.select_dtypes(include=bool).columns
df[bool_cols] = df[bool_cols].astype(int)

corr = X.corr()
sns.heatmap(corr, annot=True)

X_const = sm.add_constant(X)
vif = pd.DataFrame({
    'var': X_const.columns,
    'VIF': [variance_inflation_factor(X_const.values, i)
            for i in range(X_const.shape[1])]
})

from sklearn.linear_model import LassoCV
from sklearn.model_selection import KFold

lasso = LassoCV(cv=KFold(5, shuffle=True, random_state=42)).fit(X, y)
coef = pd.Series(lasso.coef_, index=X.columns)
print(coef)
print("Selected vars:", coef[coef != 0].index.tolist())

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.15, random_state=42
)

X_train_const = sm.add_constant(X_train, prepend=True)
X_test_const  = sm.add_constant(X_test,  prepend=True)
# Entrenamiento del modelo y visualización de los resultados iniciales
```

```
model = sm.OLS(y_train, X_train_const).fit()
print(model.summary())


y_pred = model.predict(X_test_const)
rmse = np.sqrt(mean_squared_error(y_test, y_pred))
r2_train = model.rsquared
r2_adj = model.rsquared_adj


print(f"RMSE_test: {rmse:.2f}")
print(f"R2_train:  {r2_train:.3f}")
print(f"R2_adj:    {r2_adj:.3f}")


coef_df = (
    pd.DataFrame({
        "variable": model.params.index,
        "coef": model.params.values,
        "p_value": model.pvalues.values
    })
    .set_index("variable")
    .sort_values("p_value")
)
```

Código 2: Modelo regresión linear

```
import pandas as pd
import numpy as np


from sklearn.pipeline import Pipeline
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import StandardScaler, FunctionTransformer
from sklearn.impute import SimpleImputer
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score


ruta_excel = "idealista_listings_ventas.xlsx"
```

```
# Si columna de precio tiene otro nombre, cámbiala aquí
df = pd.read_excel(ruta_excel)

# ---- 1) Definición de columnas
num_cols = ["Precio", "Superficie", "Habitaciones", "Baños", "Planta"]
bin_cols = ["Aparcamiento", "Exterior"]
cat_cols = []  # no variables categoricas

# Asegurar que las binarias son 0/1 si vienen como bool/NaN
df[bin_cols] = df[bin_cols].fillna(False).astype(int)

# ---- 2) Preprocesado
num_pipe = Pipeline([
    ("imputer", SimpleImputer(strategy="median")),
    ("scaler", StandardScaler())
])

bin_pipe = Pipeline([
    ("imputer", SimpleImputer(strategy="most_frequent")),
    ("toint", FunctionTransformer(lambda X: X.astype(int)))
])

pre = ColumnTransformer([
    ("num", num_pipe, num_cols),
    ("bin", bin_pipe, bin_cols),
], remainder="drop")

X = df[num_cols + bin_cols + cat_cols]
X_pre = pre.fit_transform(X)

# ---- 3) Elegir k por silhouette
k_candidates = range(2, 8)  # prueba 2..7 perfiles
scores = {}
for k in k_candidates:
    km = KMeans(n_clusters=k, n_init="auto", random_state=42)
```

```
    labels = km.fit_predict(X_pre)
    scores[k] = silhouette_score(X_pre, labels)

best_k = max(scores, key=scores.get)
print("Silhouette por k:", scores)
print("k óptimo:", best_k)

# ---- 4) Entrenar KMeans final y asignar perfiles
kmeans = KMeans(n_clusters=best_k, n_init="auto", random_state=42)
labels = kmeans.fit_predict(X_pre)
df["perfil"] = labels

# ---- 5) Resumen de perfiles (para “ponerles nombre”)
resumen_num = df.groupby("perfil")[num_cols].median().round(2)
resumen_bin = df.groupby("perfil")[bin_cols].mean().round(3)
print("\nMedianas numéricas por perfil:\n", resumen_num)
print("\nProporciones (binarias) por perfil:\n", resumen_bin)

# (Opcional) Etiquetas amigables según reglas simples
def bautizar_fila(row, p_med, m2_med):
    tag = []
    tag.append("caro" if row["Precio"] >= p_med else "barato")
    tag.append("grande" if row["Superficie"] >= m2_med else "pequeño")
    if row["Aparcamiento"] == 1: tag.append("con parking")
    if row["Exterior"] == 1: tag.append("exterior")
    return " · ".join(tag)
p_med_global = df["Precio"].median()
m2_med_global = df["Superficie"].median()
df["perfil_nombre"] = df.apply(
    bautizar_fila, axis=1,
    p_med=p_med_global, m2_med=m2_med_global
)

# Vista rápida del mix por perfil
```

```
print("\nEjemplos de nombres por perfil:")
print(df.groupby("perfil")["perfil_nombre"].agg(lambda s: s.value_counts().head(3)))


df["perfil_nombre"]
```

Código 3: Modelo K-Means

OLS Regression Results

Dep. Variable:	Precio (€)	R-squared:	0.787
Model:	OLS	Adj. R-squared:	0.752
Method:	Least Squares	F-statistic:	22.23
Date:	Wed, 20 Aug 2025	Prob (F-statistic):	1.44e-19
Time:	13:33:59	Log-Likelihood:	-1270.8
No. Observations:	85	AIC:	2568.
Df Residuals:	72	BIC:	2599.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.477e+05	4.65e+05	0.533	0.596	-6.79e+05	1.17e+06
Superficie (m2)	1.846e+04	1573.913	11.732	0.000	1.53e+04	2.16e+04
Habitaciones	-4.663e+05	1.18e+05	-3.951	0.000	-7.02e+05	-2.31e+05
Baños	-1.776e+05	1.41e+05	-1.262	0.211	-4.58e+05	1.03e+05
Reformado	-1.134e+05	2e+05	-0.569	0.571	-5.11e+05	2.84e+05
Terraza	-1.874e+05	2.14e+05	-0.877	0.383	-6.13e+05	2.38e+05
Planta	9.779e+04	4.66e+04	2.100	0.039	4980.719	1.91e+05
Ascensor	-7.1e+04	4.65e+05	-0.153	0.879	-9.98e+05	8.56e+05
Aparcamiento	-1.767e+05	2.45e+05	-0.721	0.473	-6.65e+05	3.12e+05
Amueblado	-5.639e+04	4.65e+05	-0.121	0.904	-9.84e+05	8.71e+05
Aire acondicionado	-2.67e+05	1.96e+05	-1.359	0.178	-6.59e+05	1.25e+05
Piscina	1.135e+05	2.76e+05	0.411	0.682	-4.37e+05	6.64e+05
Exterior	2.511e+05	2.89e+05	0.868	0.388	-3.25e+05	8.28e+05

Omnibus:	68.260	Durbin-Watson:	1.706
Prob(Omnibus):	0.000	Jarque-Bera (JB):	909.294
Skew:	2.091	Prob(JB):	3.54e-198
Kurtosis:	18.468	Cond. No.	1.32e+03

Ilustración 18: Resultados modelo propiedades en venta (datos recogidos a mano)

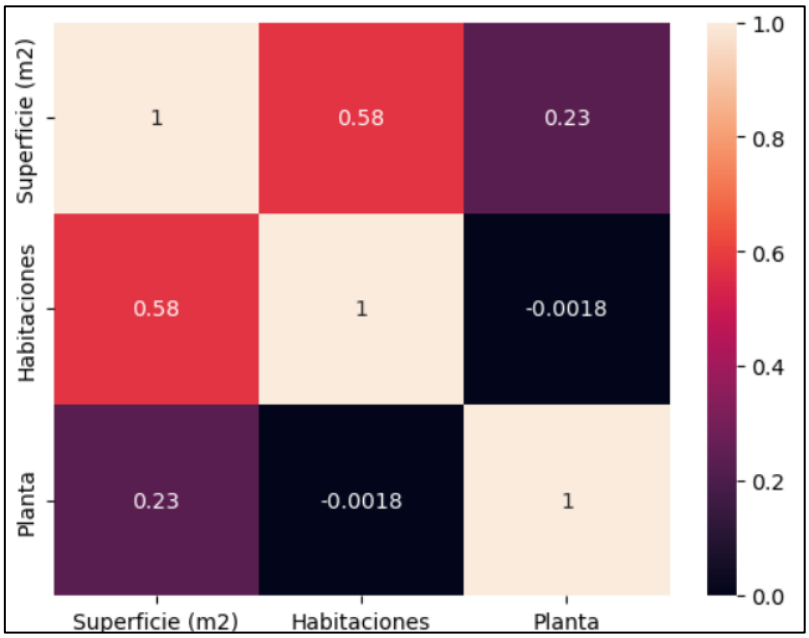


Ilustración 19: Matriz de correlación modelo final de propiedades en venta (datos recogidos a mano)

OLS Regression Results

Dep. Variable:	Precio (€)	R-squared:	0.779
Model:	OLS	Adj. R-squared:	0.721
Method:	Least Squares	F-statistic:	13.26
Date:	Wed, 20 Aug 2025	Prob (F-statistic):	4.60e-11
Time:	13:08:21	Log-Likelihood:	-872.96
No. Observations:	58	AIC:	1772.
Df Residuals:	45	BIC:	1799.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-5.24e+04	1.24e+06	-0.042	0.966	-2.54e+06	2.44e+06
Superficie (m2)	2.14e+04	2145.142	9.976	0.000	1.71e+04	2.57e+04
Habitaciones	-6.318e+05	1.63e+05	-3.868	0.000	-9.61e+05	-3.03e+05
Baños	-2.47e+05	1.83e+05	-1.349	0.184	-6.16e+05	1.22e+05
Reformado	-1.802e+05	2.79e+05	-0.646	0.521	-7.42e+05	3.81e+05
Terraza	-3.276e+05	3.01e+05	-1.090	0.281	-9.33e+05	2.78e+05
Planta	9.814e+04	5.84e+04	1.681	0.100	-1.94e+04	2.16e+05
Ascensor	3.609e+05	1.05e+06	0.343	0.733	-1.76e+06	2.48e+06
Aparcamiento	-2.724e+05	2.92e+05	-0.933	0.356	-8.6e+05	3.15e+05
Amueblado	-1.882e+05	7.42e+05	-0.254	0.801	-1.68e+06	1.31e+06
Aire acondicionado	-5.63e+05	2.88e+05	-1.953	0.057	-1.14e+06	1.75e+04
Piscina	3.896e+05	3.75e+05	1.040	0.304	-3.65e+05	1.14e+06
Exterior	5.814e+05	5.32e+05	1.092	0.281	-4.91e+05	1.65e+06

Omnibus:	26.496	Durbin-Watson:	1.707
Prob(Omnibus):	0.000	Jarque-Bera (JB):	103.913
Skew:	1.057	Prob(JB):	2.73e-23
Kurtosis:	9.207	Cond. No.	2.81e+03

Ilustración 20: Resultados modelo propiedades en venta (datos recogidos a mano) Superficie > 100 m2

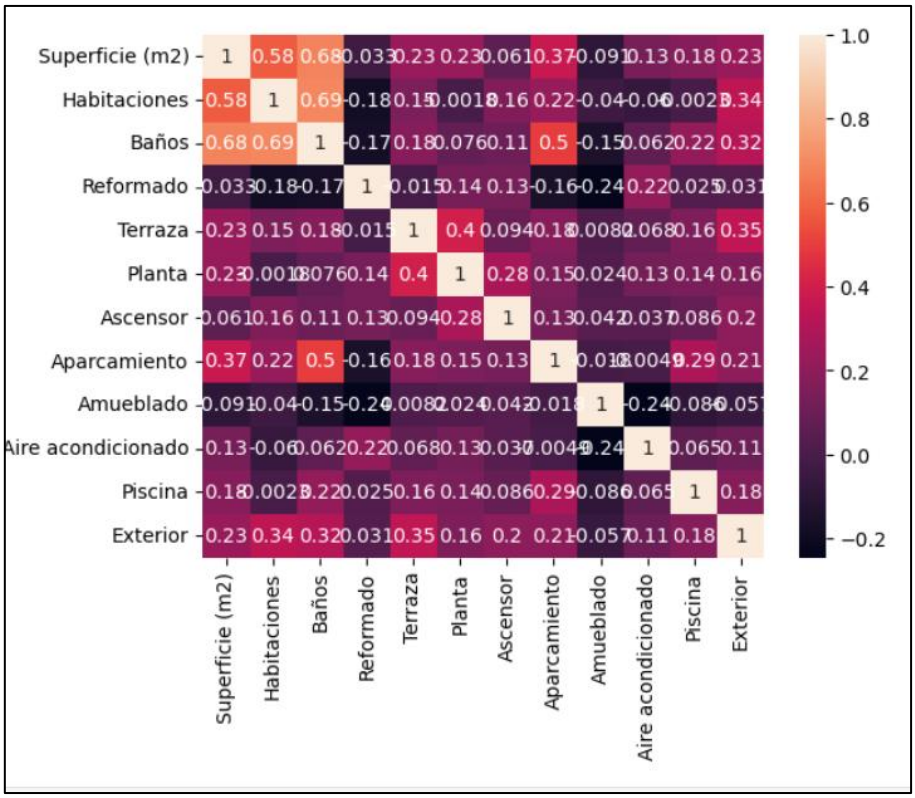


Ilustración 21: Matriz correlación modelo ventas (datos recogidos a mano) superficie > 100 m2

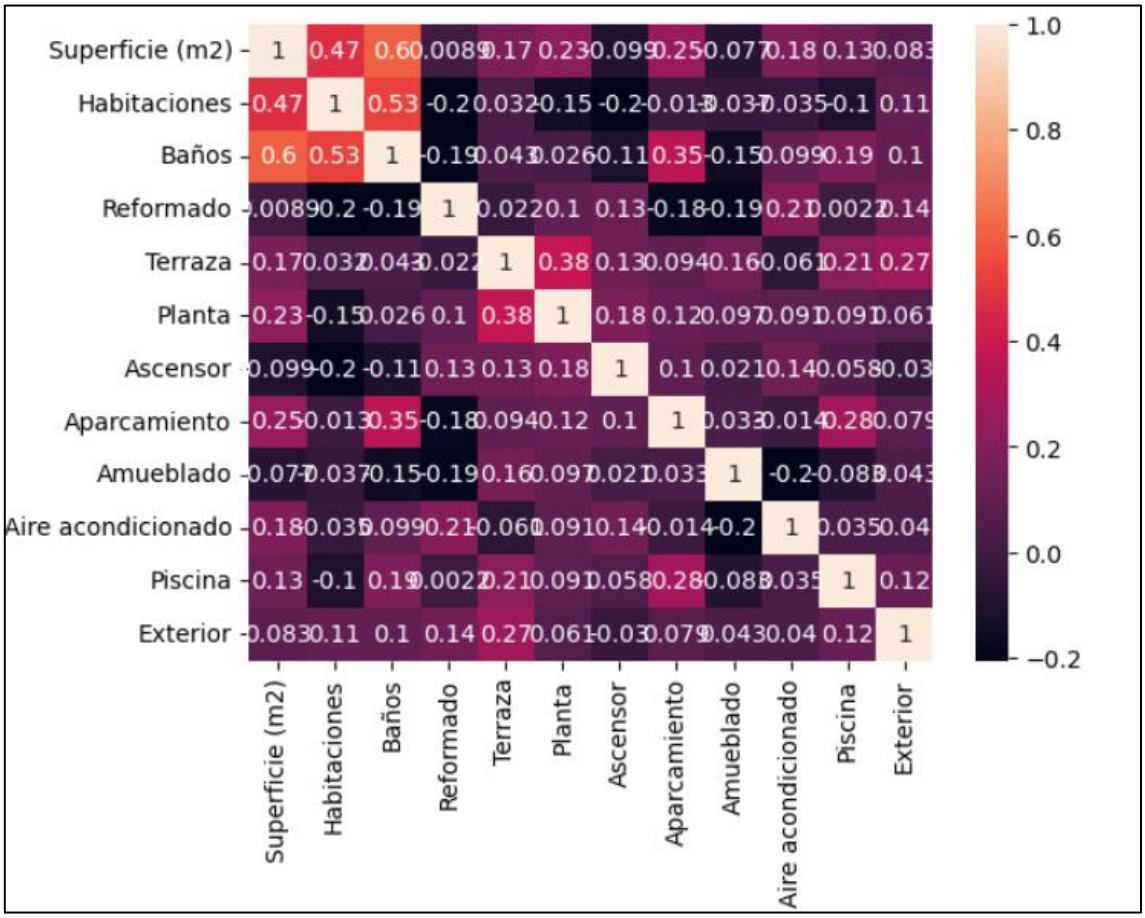


Ilustración 22: Matriz correlación modelo propiedades en venta (datos recogidos a mano) Precio > 1,000,000€

OLS Regression Results

Dep. Variable:	Precio	R-squared:	0.641
Model:	OLS	Adj. R-squared:	0.633
Method:	Least Squares	F-statistic:	77.91
Date:	Thu, 21 Aug 2025	Prob (F-statistic):	5.51e-29
Time:	18:14:37	Log-Likelihood:	-1061.8
No. Observations:	135	AIC:	2132.
Df Residuals:	131	BIC:	2143.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	777.3102	140.034	5.551	0.000	500.290	1054.330
Superficie	6.9984	1.502	4.660	0.000	4.027	9.970
Baños	493.4845	106.303	4.642	0.000	283.191	703.778
Exterior	431.4134	125.883	3.427	0.001	182.386	680.441

Omnibus:	41.959	Durbin-Watson:	1.934
Prob(Omnibus):	0.000	Jarque-Bera (JB):	98.050
Skew:	1.267	Prob(JB):	5.11e-22
Kurtosis:	6.317	Cond. No.	333.

Ilustración 23: Resultados modelo final propiedades en alquiler

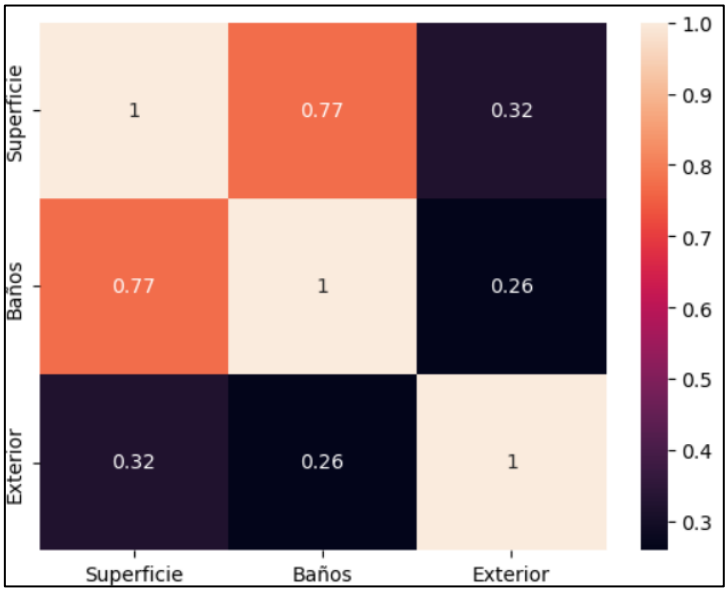


Ilustración 24: Matriz correlación modelo final propiedades en alquiler

OLS Regression Results

Dep. Variable:	Precio	R-squared:	0.977
Model:	OLS	Adj. R-squared:	0.976
Method:	Least Squares	F-statistic:	1208.
Date:	Thu, 21 Aug 2025	Prob (F-statistic):	3.73e-70
Time:	18:28:24	Log-Likelihood:	-1271.8
No. Observations:	90	AIC:	2552.
Df Residuals:	86	BIC:	2562.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-8.435e+04	8.68e+04	-0.971	0.334	-2.57e+05	8.83e+04
Superficie	1.071e+04	228.373	46.883	0.000	1.03e+04	1.12e+04
Habitaciones	-7.528e+04	3.13e+04	-2.404	0.018	-1.38e+05	-1.3e+04
Planta	3.724e+04	1.69e+04	2.207	0.030	3701.195	7.08e+04

Omnibus:	20.271	Durbin-Watson:	1.513
Prob(Omnibus):	0.000	Jarque-Bera (JB):	134.409
Skew:	0.048	Prob(JB):	6.51e-30
Kurtosis:	8.986	Cond. No.	643.

Ilustración 25: Resultados modelo final propiedades en venta (vía API)

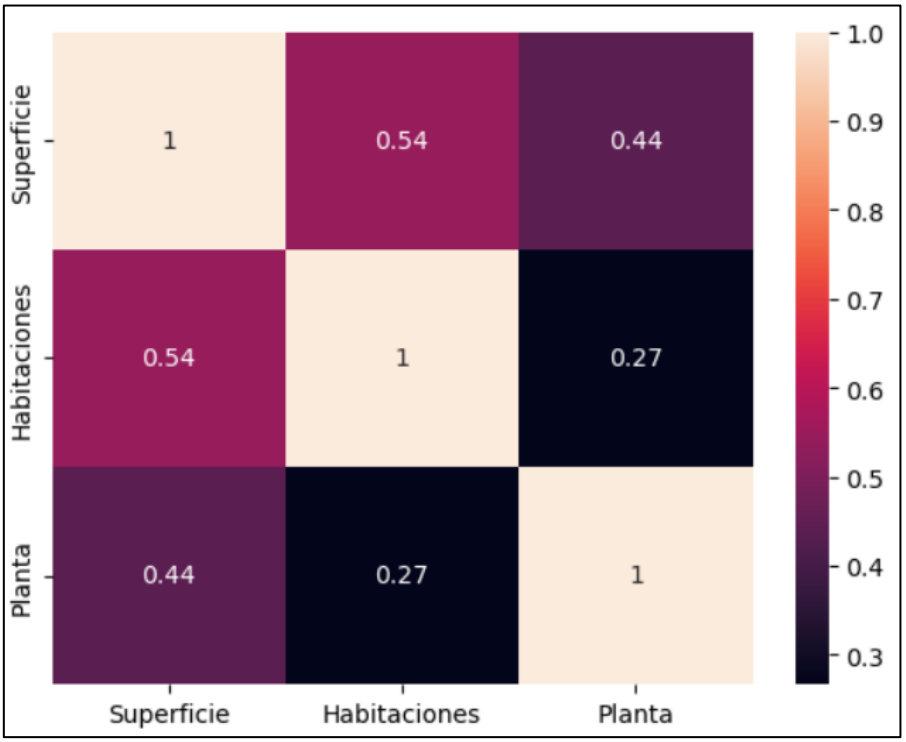


Ilustración 26: Matriz de correlación modelo final propiedades en venta (vía API)

OLS Regression Results

Dep. Variable:	Precio (€)	R-squared:	0.889
Model:	OLS	Adj. R-squared:	0.882
Method:	Least Squares	F-statistic:	134.0
Date:	Fri, 22 Aug 2025	Prob (F-statistic):	8.48e-53
Time:	11:03:01	Log-Likelihood:	-1867.7
No. Observations:	125	AIC:	3751.
Df Residuals:	117	BIC:	3774.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-8.577e+04	2.6e+05	-0.330	0.742	-6.01e+05	4.29e+05
Superficie (m2)	1.127e+04	596.479	18.899	0.000	1.01e+04	1.25e+04
Habitaciones	-1.658e+05	7.31e+04	-2.268	0.025	-3.1e+05	-2.1e+04
Baños	3.027e+04	1.02e+05	0.297	0.767	-1.72e+05	2.32e+05
Exterior	1.304e+05	2e+05	0.652	0.516	-2.66e+05	5.27e+05
Ascensor	-9.437e+04	2.24e+05	-0.421	0.675	-5.39e+05	3.5e+05
Planta	7.611e+04	3.06e+04	2.491	0.014	1.56e+04	1.37e+05
Aparcamiento	7.754e+04	1.86e+05	0.418	0.677	-2.9e+05	4.45e+05

Omnibus:	193.535	Durbin-Watson:	1.966
Prob(Omnibus):	0.000	Jarque-Bera (JB):	17326.380
Skew:	6.028	Prob(JB):	0.00
Kurtosis:	59.403	Cond. No.	1.12e+03

Ilustración 27: Resultados modelo propiedades en venta (vía API y recogidos a mano)

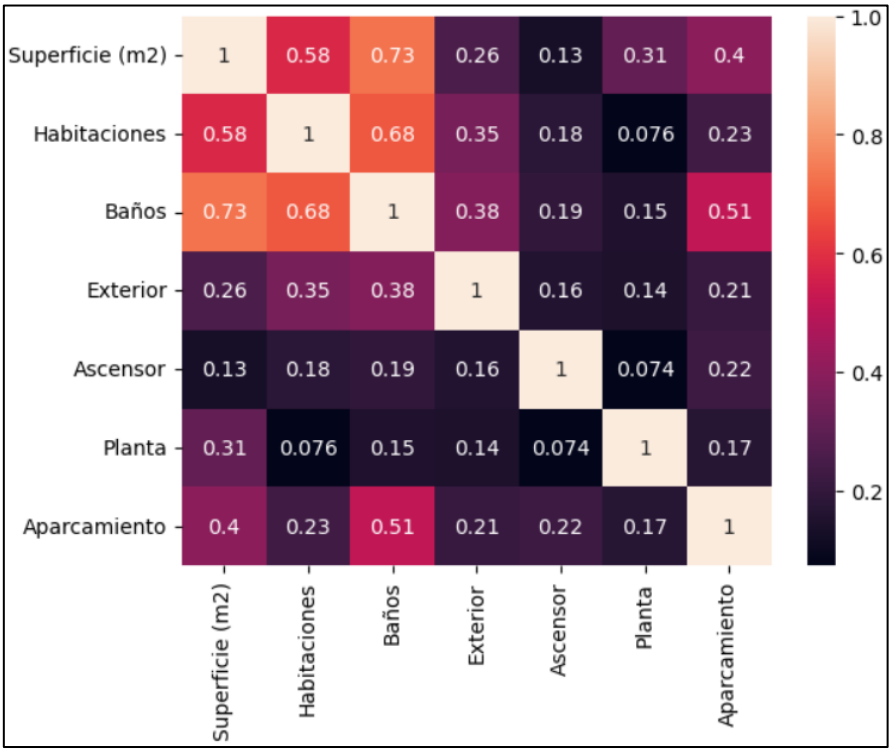


Ilustración 28: Matriz correlación modelo propiedades en venta (vía API y datos recogidos a mano)