



Facultad de Ciencias Económicas y Empresariales

ICADE

**Análisis de los determinantes del precio en el segmento de
habitaciones privadas de Airbnb Madrid: una
aproximación desde Business Analytics**

Trabajo Fin de Grado

Autor: Francisco Javier Meliá García

Director: Raúl González Fabre

MADRID | Junio 2026

Resumen

Este Trabajo Fin de Grado desarrolla un análisis cuantitativo de los determinantes del precio en el segmento de habitaciones privadas y alojamientos económicos de Airbnb en la ciudad de Madrid. El estudio se enmarca dentro del ámbito de Business Analytics y combina técnicas de análisis exploratorio de datos con modelización econométrica para identificar qué factores explican las diferencias de precio observadas entre anuncios.

Partiendo de una base de datos pública extraída de Airbnb el 15 de enero de 2021 (Messaadi, 2025), compuesta por 1.993 observaciones y 85 variables, se ha desarrollado un proceso de limpieza, depuración y construcción de variables analíticas que ha permitido obtener una muestra final de 1.438 anuncios válidos. Sobre esta muestra se ha estimado un modelo de regresión lineal múltiple mediante mínimos cuadrados ordinarios (OLS) con errores estándar robustos HC3, utilizando como variable dependiente el logaritmo del precio. Adicionalmente, se ha realizado una validación predictiva fuera de muestra mediante una partición 80/20 entre entrenamiento (1.150 observaciones) y prueba (288 observaciones).

Los resultados muestran que las variables con mayor poder explicativo son el número de huéspedes incluidos en el precio base, el nivel de equipamiento del alojamiento, el número de reseñas acumuladas, el tipo de propiedad y la localización territorial. La valoración media del alojamiento presenta un efecto negativo y estadísticamente significativo una vez controladas las demás variables. El modelo explica un 47,8 % de la variabilidad del logaritmo del precio y mantiene una capacidad predictiva razonable fuera de muestra ($R^2 = 0,450$; MAE = 0,168; RMSE = 0,223), sin evidencias de sobreajuste severo. Asimismo, el número de baños del inmueble muestra un efecto negativo sobre el precio, coherente con el carácter compartido de las habitaciones analizadas.

Palabras clave: Airbnb, pricing, precios hedónicos, regresión lineal múltiple, Business Analytics, Madrid.

Abstract

This Final Degree Project develops a quantitative analysis of the determinants of price within the private-room and budget segment of Airbnb listings in the city of Madrid. The study is framed within the Business Analytics domain and combines exploratory data analysis with econometric modelling to identify which factors explain price differences across listings.

Starting from a public dataset extracted from Airbnb on 15 January 2021 (Messaadi, 2025), comprising 1,993 observations and 85 variables, a process of data cleaning, debugging and analytical variable construction has been carried out, yielding a final sample of 1,438 valid listings. On this sample, a multiple linear regression model has been estimated by Ordinary Least Squares (OLS) with HC3 robust standard errors, using the log of price as the dependent variable. An out-of-sample predictive validation has also been performed through an 80/20 train-test split (1,150 training observations; 288 test observations).

The results show that the most explanatory variables are the number of guests included in the base price, the level of amenities offered, the cumulative number of reviews, the property type and the territorial location. The average rating of the listing has a negative and statistically significant effect once the remaining variables are controlled for. The model explains 47.8 % of the variability of the log of price and retains a reasonable predictive capacity out of sample ($R^2 = 0.450$; MAE = 0.168; RMSE = 0.223), with no evidence of severe overfitting.

Keywords: Airbnb, pricing, hedonic prices, multiple linear regression, Business Analytics, Madrid.

Índice

Resumen.....	2
Abstract	3
Índice.....	4
Índice de tablas	6
Índice de ilustraciones.....	7
Glosario.....	8
1. Introducción	10
1.1. Objetivos	10
1.2. Estado de la cuestión.....	11
1.3. Estructura del trabajo	13
2. Marco teórico	14
2.1. Plataformas digitales y mercado de Airbnb	14
2.2. Pricing y toma de decisiones.....	14
2.3. Teoría de precios hedónicos.....	15
2.4. Reputación y señales de calidad	15
2.5. Business Analytics aplicado al pricing	15
2.6. Conexión con el análisis empírico	16
3. Metodología	17
3.1. Base de datos y unidad de análisis	17
3.2. Limpieza y depuración de los datos.....	18
3.3. Construcción de variables analíticas	19
3.4. Selección de variables para el modelo	21
3.5. Especificación del modelo econométrico	24
3.6. Evaluación predictiva.....	25
4. Análisis e interpretación de resultados	27

4.1. Análisis exploratorio de la variable precio	27
4.2. Relación entre amenities y precio	28
4.3. Diferencias territoriales en el precio	30
4.4. Análisis de correlaciones	31
4.5. Resultados del modelo econométrico	32
4.6. Diagnóstico del modelo	37
4.7. Evaluación predictiva del modelo.....	40
5. Conclusiones	43
5.1. Principales hallazgos del estudio	43
5.2. Implicaciones prácticas y utilidad del modelo.....	44
5.3. Limitaciones del trabajo.....	44
5.4. Líneas futuras de investigación.....	45
6. Declaración de uso de herramientas de IAG.....	47
7. Referencias.....	49
Anexo. Código del análisis	50

Índice de tablas

Tabla 1. Variables independientes utilizadas en el modelo econométrico	21
Tabla 2. Variables descartadas y motivo de exclusión	23
Tabla 3. Partición de la muestra: entrenamiento y prueba	25
Tabla 4. Resultados del modelo de regresión lineal múltiple (variable dependiente: log_price)	34
Tabla 5. Métricas de ajuste y evaluación predictiva del modelo	42

Índice de ilustraciones

Figura 1. Distribución del precio original y del logaritmo del precio	28
Figura 2. Distribución del precio según el número de amenities (boxplot).....	29
Figura 3. Distribución de $\log(1+\text{price})$ por barrio agrupado	30
Figura 4. Matriz de correlaciones entre variables numéricas	31
Figura 5. Gráfico de coeficientes estimados con intervalos de confianza	36
Figura 6. Valores reales vs. valores ajustados por el modelo	37
Figura 7. Histograma de los residuos del modelo.....	38
Figura 8. QQ-plot de los residuos.....	39
Figura 9. Valores reales vs. valores predichos sobre el conjunto de prueba	41

Glosario

A continuación se recogen, por orden alfabético, los principales términos técnicos utilizados a lo largo del trabajo, con el objetivo de facilitar la lectura y comprensión del mismo.

Airbnb. Plataforma digital que conecta anfitriones que ofrecen alojamientos de corta estancia con viajeros que buscan alojamiento temporal.

Amenities. Conjunto de servicios o comodidades ofrecidas por un alojamiento, como wifi, cocina, aire acondicionado, calefacción o aparcamiento.

Análisis exploratorio de datos (EDA). Proceso inicial de análisis estadístico y visualización utilizado para comprender la estructura de una base de datos, detectar patrones y localizar anomalías.

Business Analytics. Disciplina orientada al uso de datos, modelos estadísticos y herramientas cuantitativas para mejorar la toma de decisiones empresariales.

Coefficiente de regresión. Valor estimado en una regresión que mide el impacto de una variable explicativa sobre la variable dependiente.

Coefficiente de correlación. Medida estadística, expresada mediante el coeficiente de correlación, que indica la intensidad y la dirección de la relación lineal entre dos variables.

Dataset. Conjunto estructurado de datos utilizado para el análisis.

Dummy variable. Variable binaria (0/1) utilizada para representar categorías cualitativas, como barrios o tipos de alojamiento.

Errores robustos HC3. Ajuste estadístico de los errores estándar que corrige posibles problemas de heterocedasticidad y mejora la inferencia del modelo.

Heterocedasticidad. Situación en la que la varianza de los errores no es constante entre observaciones.

Histograma. Gráfico que representa la distribución de frecuencias de una variable numérica.

Listing. Anuncio individual publicado en Airbnb correspondiente a un alojamiento.

$\log(1+\text{price}) / \log_price$. Transformación matemática del precio definida como $\log(1+\text{precio})$, utilizada para mejorar el ajuste estadístico del modelo y facilitar la interpretación porcentual.

Machine Learning. Conjunto de técnicas algorítmicas orientadas a detectar patrones y realizar predicciones a partir de datos.

MAE (Mean Absolute Error). Error absoluto medio entre los valores reales y los predichos por el modelo.

Mínimos Cuadrados Ordinarios (OLS). Método econométrico utilizado para estimar modelos de regresión lineal.

Modelo econométrico. Representación matemática utilizada para explicar relaciones entre variables económicas o empresariales.

Pricing. Estrategia de fijación de precios de un producto o servicio.

Precios hedónicos. Enfoque teórico según el cual el precio de un bien depende del valor implícito de sus atributos.

R² (coeficiente de determinación). Medida que indica qué proporción de la variabilidad de la variable dependiente explica el modelo.

Revenue Management. Técnica empresarial orientada a maximizar ingresos ajustando precios según demanda y capacidad disponible.

RMSE (Root Mean Squared Error). Raíz del error cuadrático medio; penaliza más los errores grandes que el MAE.

Significatividad estadística. Evidencia de que una relación observada entre variables difícilmente se debe al azar.

Train-test split. Técnica que divide la muestra en un conjunto de entrenamiento (para estimar el modelo) y otro de prueba (para evaluar su capacidad predictiva fuera de muestra).

Variable dependiente. Variable que se desea explicar; en este TFG, el precio del alojamiento.

Variable explicativa. Variable utilizada para explicar cambios en la variable dependiente.

Visualización de datos. Representación gráfica de información estadística para facilitar su interpretación.

1. Introducción

La digitalización ha transformado de manera profunda numerosos sectores económicos, facilitando la aparición de plataformas capaces de conectar oferta y demanda de forma más eficiente y escalable. En el ámbito del turismo, Airbnb se ha consolidado como uno de los ejemplos más representativos de esta transformación, al permitir que particulares y empresas comercialicen alojamientos de corta duración a través de un entorno digital global.

En ciudades con elevada demanda turística como Madrid, la expansión de Airbnb ha modificado tanto los hábitos de consumo como la estructura competitiva del mercado del alojamiento. A diferencia del sector hotelero tradicional, donde los productos presentan un mayor grado de estandarización, en Airbnb cada alojamiento constituye una oferta diferenciada en términos de localización, capacidad, características físicas, servicios ofrecidos y reputación online. Esta heterogeneidad genera un entorno competitivo en el que la fijación de precios resulta especialmente compleja.

Desde una perspectiva económica, el precio de los alojamientos en plataformas digitales no depende únicamente de factores tradicionales como la localización o el tamaño, sino también de elementos específicos del entorno digital, como las valoraciones de los usuarios, el número de reseñas o la calidad percibida del anuncio. En consecuencia, el análisis de los determinantes del precio requiere un enfoque que combine teoría económica y técnicas cuantitativas aplicadas a datos reales.

En este contexto, el análisis del precio en Airbnb constituye un problema especialmente relevante dentro del ámbito de *Business Analytics*. El uso de herramientas de análisis de datos permite identificar patrones, estimar relaciones entre variables y generar conocimiento útil para la toma de decisiones. Comprender qué factores están asociados a precios más elevados no solo tiene interés académico, sino que también puede aportar valor práctico para los anfitriones, al facilitar la optimización de sus estrategias de *pricing*, así como para analistas y gestores del sector turístico.

Por tanto, el presente trabajo se centra en el estudio del segmento de habitaciones privadas y alojamientos económicos anunciados en Airbnb en la ciudad de Madrid, con el objetivo de contribuir a una mejor comprensión de los factores que influyen en la fijación de precios en plataformas digitales de alojamiento.

1.1. Objetivos

El objetivo general de este trabajo es analizar los principales factores que influyen en el precio de las habitaciones privadas y alojamientos de rango económico e intermedio anunciados en Airbnb en la ciudad de Madrid, mediante el uso de técnicas de análisis de datos y modelización cuantitativa aplicadas a una base de datos de anuncios extraídos de la plataforma en enero de 2021.

A partir de este objetivo general, se plantean los siguientes objetivos específicos:

- Analizar la distribución del precio de los alojamientos incluidos en la muestra y evaluar la conveniencia de aplicar transformaciones estadísticas sobre la variable dependiente para mejorar el ajuste del modelo.
- Identificar y seleccionar las variables relevantes que pueden influir en el precio, distinguiendo entre atributos físicos del alojamiento (como número de camas y baños), indicadores de calidad y reputación (valoraciones y número de reseñas) y servicios ofrecidos.
- Evaluar la existencia de diferencias de precio entre barrios de Madrid, considerando la localización como uno de los principales determinantes del valor dentro del mercado de alojamiento turístico.
- Estimar un modelo de regresión lineal múltiple que permita cuantificar el efecto de las variables explicativas sobre el precio de los alojamientos analizados.
- Analizar la capacidad predictiva del modelo mediante técnicas de validación fuera de muestra, con el fin de evaluar la capacidad del modelo para aproximar el comportamiento de nuevos datos.
- Interpretar los resultados obtenidos desde una perspectiva económica y de negocio, extrayendo implicaciones útiles para la comprensión de las dinámicas de *pricing* dentro del segmento analizado de Airbnb Madrid.

1.2. Estado de la cuestión

La literatura académica sobre Airbnb ha experimentado un notable crecimiento en los últimos años, abordando tanto su impacto en el mercado turístico e inmobiliario como el análisis de los determinantes del precio de los alojamientos. Dentro de esta segunda línea, numerosos estudios han aplicado modelos de precios hedónicos con el objetivo de identificar qué factores explican la variabilidad de precios entre anuncios.

Desde un punto de vista teórico, la mayor parte de estos trabajos se fundamenta en el modelo de precios hedónicos desarrollado por Rosen (1974), según el cual el precio de un bien heterogéneo puede expresarse como una función de sus atributos. Este enfoque resulta especialmente adecuado para el análisis de Airbnb, donde cada alojamiento presenta una combinación única de características físicas, localización, servicios ofrecidos y reputación online.

En el ámbito empírico, Wang y Nicolau (2017) analizan los determinantes del precio en distintas ciudades mediante modelos de regresión hedónica aplicados a datos de Airbnb. Su estudio muestra que la localización constituye uno de los factores más relevantes, observándose primas significativas en zonas céntricas y turísticas. Asimismo, variables relacionadas con el

tamaño del alojamiento presentan un efecto positivo sobre el precio, lo que refleja la mayor utilidad que ofrecen al consumidor.

Por su parte, Gibbs et al. (2018) examinan varias ciudades utilizando modelos econométricos que incorporan tanto variables estructurales como reputacionales. Los autores encuentran que características físicas como el número de camas influyen de forma significativa en el precio, pero también destacan el papel de las valoraciones y el número de reseñas como señales de calidad que permiten a los anfitriones fijar precios más elevados.

En esta misma línea, Zervas, Proserpio y Byers (2017) analizan el impacto de las reseñas en el mercado de Airbnb y concluyen que la reputación online reduce la incertidumbre del consumidor, lo que se traduce en una mayor disposición a pagar por alojamientos mejor valorados. Este resultado pone de manifiesto la importancia de los mecanismos de información en plataformas digitales.

Asimismo, Ert, Fleischer y Magen (2016) estudian el papel de la confianza en la economía colaborativa mediante el análisis de datos de Airbnb. Sus resultados muestran que la percepción de confianza, medida a través de valoraciones y perfiles de anfitriones, tiene un efecto significativo sobre la decisión de reserva y, por tanto, sobre el precio que los usuarios están dispuestos a pagar.

En cuanto a la estructura del mercado, Guttentag (2015) analiza el crecimiento de Airbnb como una innovación disruptiva dentro del sector turístico. Su trabajo pone de relieve la elevada heterogeneidad de la oferta y la descentralización de la fijación de precios, lo que refuerza la necesidad de analizar empíricamente los factores que influyen en el precio.

Por otro lado, Dogru, Mody y Suess (2019) estudian el impacto de Airbnb en distintos mercados turísticos y analizan cómo variables como el tipo de propiedad, la localización y los servicios ofrecidos influyen en el precio. Sus resultados confirman que los alojamientos completos y con mayor nivel de equipamiento presentan precios más elevados, lo que sugiere que el valor percibido por el consumidor es un determinante clave.

Finalmente, Quattrone et al. (2016) examinan la distribución espacial de los alojamientos en Airbnb y muestran que existen diferencias significativas entre zonas, lo que refuerza la importancia de la localización como determinante del precio.

Desde el punto de vista metodológico, la mayoría de estos trabajos emplea modelos de regresión lineal múltiple por su capacidad para estimar efectos marginales de forma interpretable. No obstante, algunos estudios recientes han comenzado a incorporar técnicas de aprendizaje automático para mejorar la capacidad predictiva, aunque a menudo a costa de una menor interpretabilidad.

En conjunto, la evidencia empírica sugiere que el precio de los alojamientos en Airbnb depende de una combinación de factores estructurales (localización y tamaño), factores de calidad

(servicios ofrecidos) y factores reputacionales (valoraciones y reseñas). Sin embargo, gran parte de estos estudios se centra en ciudades como Nueva York o Barcelona, mientras que existe una menor evidencia aplicada al caso de Madrid.

En este contexto, el presente trabajo contribuye a la literatura existente mediante la aplicación de técnicas de *Business Analytics* a una base de datos real de alojamientos en Madrid, con el objetivo de identificar los principales determinantes del precio y aportar conclusiones con utilidad práctica.

Asimismo, debe tenerse en cuenta que distintos segmentos del mercado Airbnb pueden presentar dinámicas de precios diferentes. En particular, las habitaciones privadas y los alojamientos económicos pueden mostrar estructuras de precio distintas a las observadas en viviendas completas o alojamientos premium.

1.3. Estructura del trabajo

El presente Trabajo Fin de Grado se organiza en cinco capítulos principales, estructurados de forma secuencial para combinar el desarrollo teórico con el análisis empírico, siguiendo un enfoque propio de *Business Analytics*.

El Capítulo 1 introduce el contexto del estudio, justificando la relevancia del análisis del precio en plataformas digitales de alojamiento como Airbnb. Asimismo, se presentan los objetivos del trabajo y se incluye una revisión de la literatura existente sobre los determinantes del precio, que permite situar la investigación dentro del marco académico.

El Capítulo 2 desarrolla el marco teórico. En este capítulo se abordan los principales conceptos necesarios para el análisis, incluyendo el funcionamiento de las plataformas digitales, la fijación de precios en mercados turísticos y la teoría de precios hedónicos. Asimismo, se integran las aportaciones de la literatura previa, evitando duplicidades con el apartado anterior y proporcionando una base conceptual coherente para el análisis empírico.

El Capítulo 3 recoge el estudio empírico y el análisis de datos. En él se describe la base de datos utilizada, el proceso de limpieza y preparación de los datos, así como la construcción de variables. A continuación, se desarrolla el análisis exploratorio y se presenta la especificación del modelo de regresión lineal múltiple empleado para estimar los determinantes del precio.

El Capítulo 4 presenta los resultados del modelo estimado. En este capítulo se analizan la significatividad y magnitud de los coeficientes, así como la capacidad explicativa del modelo. Además, se incorpora una evaluación de su capacidad predictiva mediante técnicas de validación fuera de muestra, con el fin de comprobar su utilidad en la predicción de precios. Asimismo, se utilizan representaciones gráficas para facilitar la interpretación de los resultados.

El Capítulo 5 recoge las conclusiones del trabajo. En él se sintetizan los principales hallazgos, se discuten sus implicaciones desde una perspectiva de negocio y se señalan las limitaciones del análisis, así como posibles líneas de investigación futura.

2. Marco teórico

El análisis del precio de los alojamientos en Airbnb requiere integrar distintos enfoques teóricos relacionados con la economía digital, la fijación de precios y la analítica de datos. A diferencia del mercado hotelero tradicional, donde los productos presentan cierto grado de estandarización, Airbnb configura un mercado altamente heterogéneo en el que cada alojamiento constituye una combinación única de atributos.

En este contexto, el precio puede entenderse como el resultado de la interacción entre características físicas del alojamiento, elementos de calidad percibida, factores reputacionales y variables de localización. Este marco permite justificar el uso de modelos cuantitativos orientados a identificar los determinantes del precio.

2.1. Plataformas digitales y mercado de Airbnb

Las plataformas digitales actúan como intermediarios que conectan distintos grupos de usuarios, facilitando transacciones y reduciendo costes de búsqueda. En el caso de Airbnb, la plataforma conecta anfitriones y huéspedes en un mercado descentralizado donde cada oferente fija su propio precio.

Este tipo de mercado se caracteriza por una elevada heterogeneidad en la oferta y por la presencia de información imperfecta. Los usuarios no pueden observar completamente la calidad del alojamiento antes de la compra, lo que da lugar a la aparición de mecanismos de señalización, como las valoraciones y las reseñas.

Desde una perspectiva económica, Airbnb puede interpretarse como un mercado competitivo en el que los precios emergen de la interacción entre oferta y demanda, condicionados por la información disponible y por la comparación con alojamientos similares.

2.2. Pricing y toma de decisiones

El precio constituye una de las variables estratégicas más relevantes en cualquier mercado, ya que influye directamente en los ingresos, la demanda y el posicionamiento competitivo.

En el caso del alojamiento turístico y habitaciones privadas, la fijación de precios presenta una complejidad adicional debido a la existencia de múltiples factores que afectan a la disposición a pagar del consumidor. Entre ellos destacan la localización, la capacidad del alojamiento, el nivel de servicios ofrecidos y la calidad percibida.

Aunque en entornos profesionales se utilizan técnicas avanzadas de *revenue management*, en el caso de Airbnb muchos anfitriones fijan sus precios de forma descentralizada, basándose en la observación del mercado y en la comparación con anuncios similares. Esto refuerza la importancia de analizar empíricamente qué factores explican las diferencias de precio.

2.3. Teoría de precios hedónicos

La teoría de precios hedónicos, desarrollada por Rosen (1974), constituye el principal marco teórico para analizar bienes heterogéneos. Según este enfoque, el precio de un bien puede expresarse como una función de sus atributos individuales.

Aplicado al caso de Airbnb, esto implica que el precio de un alojamiento depende de características como:

- La localización (barrio).
- La capacidad y tamaño del alojamiento (camas, baños).
- Los servicios ofrecidos (*amenities*).
- La reputación online (valoraciones y reseñas).
- El tipo de propiedad.

Este enfoque resulta especialmente adecuado en mercados donde los productos no son homogéneos, como ocurre en Airbnb, donde cada anuncio presenta una combinación única de atributos.

Además, la teoría hedónica permite interpretar los coeficientes del modelo econométrico como el valor marginal de cada característica, lo que facilita la traducción de los resultados a implicaciones económicas.

2.4. Reputación y señales de calidad

En los mercados digitales, la información asimétrica entre oferentes y demandantes genera incertidumbre sobre la calidad del producto. Para reducir esta incertidumbre, las plataformas incorporan mecanismos de reputación que actúan como señales de calidad.

En Airbnb, las valoraciones y el número de reseñas permiten a los usuarios inferir la calidad del alojamiento a partir de la experiencia de otros consumidores. Desde un punto de vista económico, estas variables pueden influir en la disposición a pagar, lo que justifica su inclusión en el análisis del precio.

2.5. Business Analytics aplicado al pricing

El análisis del precio en Airbnb puede enmarcarse dentro del ámbito de *Business Analytics*, entendido como el uso de datos y técnicas cuantitativas para mejorar la toma de decisiones.

Dentro de este enfoque, es posible distinguir entre tres niveles de análisis:

- Analítica descriptiva, que permite comprender la estructura del mercado mediante estadísticas y visualizaciones.

- Analítica explicativa, que identifica relaciones entre variables mediante modelos econométricos.
- Analítica predictiva, que evalúa la capacidad del modelo para anticipar resultados en nuevos datos.

El presente trabajo se sitúa principalmente en el ámbito explicativo, al estimar los determinantes del precio mediante regresión lineal, pero incorpora también una dimensión predictiva mediante la validación fuera de muestra.

2.6. Conexión con el análisis empírico

Los conceptos desarrollados en este capítulo permiten justificar el enfoque metodológico adoptado en el trabajo. En particular, la teoría de precios hedónicos fundamenta la especificación del modelo, mientras que el enfoque de *Business Analytics* justifica el uso de datos reales y técnicas cuantitativas para extraer conclusiones aplicables.

En consecuencia, el siguiente capítulo desarrolla el análisis empírico del mercado de Airbnb en Madrid, aplicando las herramientas descritas para identificar los factores que influyen en la fijación de precios.

3. Metodología

Una vez desarrollado el marco teórico y revisada la literatura académica sobre los determinantes del precio en Airbnb, este capítulo presenta la metodología empleada para llevar a cabo el análisis empírico. El objetivo es explicar de forma ordenada cómo se ha pasado de la base de datos original al dataset final utilizado en el modelo, qué transformaciones se han realizado, qué variables se han seleccionado y qué técnica estadística se ha utilizado para estimar los determinantes del precio.

La metodología seguida responde a un enfoque cuantitativo propio de *Business Analytics*. En primer lugar, se parte de datos reales procedentes de anuncios de Airbnb Madrid en enero de 2021. En segundo lugar, se realiza un proceso de limpieza y preparación de la información para garantizar que los datos utilizados sean coherentes y adecuados para el análisis. En tercer lugar, se construyen variables analíticas relevantes y se seleccionan aquellas que tienen mayor sentido estadístico y económico. Finalmente, se estima un modelo de regresión lineal múltiple y se evalúa también su capacidad predictiva mediante una partición de la muestra en entrenamiento y prueba.

El propósito del modelo no es únicamente obtener una predicción exacta del precio de cada alojamiento individual, sino identificar qué factores están asociados con precios más altos o bajos. Por tanto, el análisis combina una dimensión explicativa, centrada en interpretar el efecto de las variables, con una dimensión predictiva, orientada a comprobar si el modelo es capaz de generalizar razonablemente sobre datos no utilizados durante su estimación.

3.1. Base de datos y unidad de análisis

La base de datos utilizada en este trabajo procede de anuncios publicados en Airbnb correspondientes a la ciudad de Madrid. Los datos fueron extraídos de la plataforma el 15 de enero de 2021, por lo que el análisis se basa en una fotografía transversal (*cross-sectional snapshot*) del mercado en un momento temporal concreto. En consecuencia, los resultados obtenidos deben interpretarse en el contexto específico del mercado Airbnb madrileño en dicho periodo.

La muestra analizada se concentra principalmente en habitaciones privadas y alojamientos de rango económico e intermedio, con precios comprendidos entre 10 € y 88 € por noche. Por tanto, el análisis desarrollado no representa la totalidad del mercado Airbnb en Madrid, sino el comportamiento de un segmento específico de la plataforma caracterizado por alojamientos de menor precio relativo.

Cada observación representa un alojamiento individual anunciado en Airbnb, por lo que la unidad de análisis es el anuncio o *listing*. Esto implica que el estudio no analiza reservas efectivamente realizadas, usuarios individuales ni transacciones observadas, sino las

características visibles de los alojamientos ofertados en la plataforma en el momento de extracción de los datos.

El archivo original utilizado en el análisis, *madrid_.csv*, obtenido a través de la plataforma Kaggle (Messaadi, 2025), contenía inicialmente 1.993 observaciones y 85 variables. Estas variables recogían información diversa sobre los alojamientos, incluyendo precio, localización, características físicas, capacidad, servicios ofrecidos, reputación online, tipo de propiedad y atributos descriptivos del anuncio.

No obstante, no todas las variables del *dataset* original resultaban adecuadas para el modelo econométrico: algunas presentaban un elevado número de valores ausentes, otras aportaban información poco relevante o redundante y varias requerían una transformación previa. La selección de las variables explicativas se apoyó, además, en un análisis de correlaciones entre las variables numéricas del conjunto original y el precio (recogido en el anexo), que permitió priorizar las más asociadas al precio y descartar las redundantes o de escasa relación.

Tras la limpieza, depuración y selección de variables se obtuvo una muestra final de 1.438 observaciones. El modelo se construye sobre 8 variables explicativas (que, al codificar las categóricas tipo de propiedad y barrio mediante variables dummy, se traducen en 18 regresores), además de la variable dependiente. La reducción respecto al archivo original se debió sobre todo a la eliminación de registros incompletos en variables clave del modelo y a la exclusión de variables redundantes o poco informativas.

Esta muestra final se considera suficiente para los objetivos del estudio: el modelo se estimó sobre 1.150 observaciones de entrenamiento (80 %) y se validó sobre 288 de prueba (20 %), manteniendo una relación holgada entre el número de observaciones y de variables explicativas que favorece la estabilidad de la estimación y la interpretabilidad de los resultados.

3.2. Limpieza y depuración de los datos

La limpieza de datos constituye una fase esencial en cualquier proyecto de *Business Analytics*. Los datos reales procedentes de plataformas digitales suelen presentar problemas de formato, información incompleta, variables no estructuradas y posibles inconsistencias. Si estos problemas no se corrigen antes de estimar el modelo, los resultados pueden ser poco fiables o incluso incorrectos.

En primer lugar, se revisó la estructura general del dataset original. Para ello, se analizaron los nombres de las variables, el tipo de dato asociado a cada columna y la presencia de valores perdidos. Esta revisión permitió distinguir entre variables potencialmente relevantes para explicar el precio y variables que no aportaban información útil para el objetivo del trabajo.

En segundo lugar, se eliminaron variables con escasa utilidad analítica. Algunas columnas podían contener información descriptiva del anuncio, identificadores internos o variables con demasiados valores ausentes. Este tipo de información no resulta adecuada para un modelo

econométrico, ya que puede introducir ruido, reducir el tamaño muestral disponible o dificultar la interpretación de los resultados.

En tercer lugar, se revisaron los valores faltantes en las variables relevantes. Para la estimación del modelo se trabajó únicamente con observaciones completas en las variables seleccionadas. Esta decisión implica una reducción del tamaño de la muestra, pero permite garantizar que todos los registros utilizados en el modelo contengan información comparable.

También fue necesario revisar el formato de determinadas variables. En bases de datos procedentes de plataformas digitales es habitual que variables aparentemente numéricas aparezcan almacenadas como texto. Si estas variables no se transforman correctamente, el modelo no puede utilizarlas de forma adecuada. Por ello, se convirtieron a formato numérico aquellas variables que representaban cantidades, como el precio, el número de camas, la capacidad o determinadas puntuaciones.

Un ejemplo importante es la variable *guests_included*, que puede aparecer originalmente en formato textual. Para poder incorporarla al modelo, fue necesario convertirla en una variable numérica que indicase el número de huéspedes incluidos en el precio base. Esta transformación resulta relevante porque permite interpretar la variable como una medida de capacidad comercial del alojamiento.

Además, se revisaron posibles problemas derivados de columnas duplicadas o inconsistencias en nombres de variables. Este tipo de errores puede surgir durante procesos de importación o renombrado de columnas y conviene corregirlo antes del análisis para evitar problemas en la estimación posterior.

Como resultado de esta fase, se obtuvo una base de datos más limpia, coherente y preparada para la construcción de variables analíticas y la estimación del modelo.

3.3. Construcción de variables analíticas

Una vez depurada la base de datos, se procedió a construir variables con mayor valor analítico. Esta fase es importante porque muchas veces la información original no se encuentra directamente en el formato más adecuado para el análisis estadístico.

La transformación más relevante fue la creación de la variable *log_price*, definida como $\log(1 + \text{precio})$ (transformación implementada en Python mediante *np.log1p*). El precio original de los alojamientos suele presentar una distribución asimétrica hacia la derecha: la mayoría de los anuncios se concentran en precios bajos o medios, mientras que unos pocos presentan precios muy elevados. Esta distribución puede generar problemas en un modelo lineal, ya que los valores extremos tienen una influencia excesiva sobre la estimación.

El uso del logaritmo del precio permite reducir el efecto de estos valores extremos y aproximar la distribución a una forma más equilibrada. Además, esta transformación facilita la interpretación económica de los coeficientes. En un modelo donde la variable dependiente está

expresada en logaritmos, los coeficientes de las variables explicativas pueden interpretarse aproximadamente como variaciones porcentuales en el precio. Dado que en la muestra todos los precios son superiores o iguales a 10 €, la diferencia numérica entre $\log(\text{precio})$ y $\log(1 + \text{precio})$ resulta despreciable a efectos prácticos.

Otra variable construida fue *amenities_count*. La variable original *amenities* contenía, almacenada como texto, la lista de servicios ofrecidos por cada alojamiento, por lo que no podía introducirse directamente en el modelo. Para resolverlo, se transformó en una variable numérica que cuenta el número de servicios listados en cada anuncio. El conjunto de servicios registrados en la base de datos está formado por quince categorías que, por orden de frecuencia, son: *wifi*, cocina, lavadora, calefacción, ascensor, aire acondicionado, admite mascotas, secadora, aparcamiento gratuito, lavavajillas, desayuno, *self check-in*, primera línea de playa, chimenea y jacuzzi.

Esta variable funciona como un indicador sintético del nivel de equipamiento del alojamiento. Aunque no distingue entre tipos concretos de servicios, permite capturar de forma sencilla si un alojamiento cubre o no las prestaciones básicas que un huésped espera encontrar. Su inclusión en el modelo responde precisamente a este interés: contrastar en qué medida disponer de un equipamiento más completo, frente a alojamientos con lo mínimo, se traduce en un mayor precio. Desde el punto de vista económico, es razonable esperar esta relación positiva, ya que un mayor número de servicios incrementa el valor percibido por el consumidor.

Conviene, no obstante, precisar el alcance de la variable para interpretarla correctamente. El campo está limitado a un máximo de cuatro servicios por anuncio, de modo que *amenities_count* toma valores comprendidos entre uno y cuatro (media de 3,47); por ello capta sobre todo la diferencia entre los alojamientos más básicos y aquellos que ofrecen un equipamiento más completo, más que matices finos entre niveles intermedios. Además, debe tenerse en cuenta que la información es autodeclarada por los anfitriones y no está validada, por lo que puede incluir alguna etiqueta poco fiable, y que corresponde a enero de 2021: es razonable suponer que servicios entonces diferenciadores, como el wifi, tengan hoy una menor capacidad de influir en el precio al haberse generalizado su disponibilidad.

También se trabajó con variables de localización. La variable *neighbourhood_cleansed* recoge el barrio en el que se encuentra cada alojamiento. Dado que la localización es uno de los determinantes más importantes del precio en mercados inmobiliarios y turísticos, esta variable resulta fundamental para el modelo. Sin embargo, algunos barrios podían tener pocas observaciones, lo que dificultaría estimar efectos individuales de forma estable. Por ello, se agruparon determinadas categorías con baja representación, conservando los diez barrios con mayor frecuencia y agrupando el resto bajo la categoría *Otros*.

Esta decisión permite mantener la información territorial relevante sin generar un número excesivo de variables dummy. Además, mejora la estabilidad del modelo, ya que evita estimar coeficientes a partir de muy pocas observaciones.

3.4. Selección de variables para el modelo

La selección de variables se realizó combinando criterios estadísticos y criterios económicos. Este enfoque es importante porque un modelo no debe construirse únicamente a partir de correlaciones, pero tampoco debe ignorar la evidencia empírica observada en los datos.

Desde el punto de vista estadístico, se analizaron las relaciones entre las variables numéricas y la variable dependiente *log_price*. El análisis de correlaciones permitió identificar qué variables presentaban una asociación más clara con el precio y cuáles podían resultar menos relevantes. También permitió detectar posibles relaciones elevadas entre variables explicativas, lo que podría generar problemas de multicolinealidad.

La multicolinealidad, que surge cuando dos o más variables explicativas están muy correlacionadas, dificulta separar el efecto individual de cada una, por lo que se evitó incluir variables que midieran esencialmente lo mismo.

Por este motivo, la selección final de variables no se basó únicamente en la correlación, sino también en la interpretación económica de cada variable. El objetivo era construir un modelo que fuera estadísticamente razonable y, al mismo tiempo, defendible desde el punto de vista del negocio.

Las variables finalmente seleccionadas, detalladas en la Tabla 1, abarcan la capacidad y características físicas del alojamiento (*beds*, *guests_included*, *bathrooms*), su equipamiento (*amenities_count*), su reputación (*review_scores_rating*, *number_of_reviews*), el tipo de propiedad agrupado en tres categorías (apartamento, casa y otros tipos, en lugar de tratarlo como una variable continua) y la localización mediante variables dummy de barrio.

Tabla 1. Variables independientes utilizadas en el modelo econométrico

Tipo de variable	Variable	Definición	Tipo de dato	Justificación económica
Capacidad y características físicas	<i>beds</i>	Número de camas disponibles en el alojamiento.	Numérica	Aproxima la capacidad física del alojamiento. Alojamientos con mayor capacidad pueden justificar precios superiores.

Tipo de variable	Variable	Definición	Tipo de dato	Justificación económica
Capacidad y características físicas	<i>guests_included</i>	Número de huéspedes incluidos en el precio base del anuncio (puede haber 2 camas pero que el precio base contemple un solo huésped).	Numérica	Refleja una dimensión comercial de la capacidad del alojamiento y la política de precios del anfitrión.
Capacidad y características físicas	<i>bathrooms</i>	Número de baños del inmueble en el que se ubica el alojamiento.	Numérica	En una muestra de habitaciones privadas, un mayor número de baños señala una vivienda más grande y compartida, asociada a un precio inferior de la habitación.
Calidad y equipamiento	<i>amenities_count</i>	Número total de servicios o comodidades (<i>amenities</i>) ofrecidos por el alojamiento.	Numérica	Aproxima el nivel de equipamiento y el valor percibido por el huésped.
Reputación y calidad percibida	<i>review_scores_rating</i>	Valoración media otorgada por los usuarios al alojamiento.	Numérica	Actúa como señal de calidad y confianza dentro de la plataforma.
Reputación y calidad percibida	<i>number_of_reviews</i>	Número acumulado de reseñas recibidas por el alojamiento.	Numérica	Aproxima la experiencia y visibilidad del anuncio dentro de Airbnb.
Tipo de propiedad	<i>property_type_grouped</i>	Tipo de propiedad agrupado en tres categorías: apartamento, casa y otros tipos.	Categórica (dummy)	Permite controlar diferencias estructurales entre distintos tipos de alojamiento.

Tipo de variable	Variable	Definición	Tipo de dato	Justificación económica
Localización	<i>neighbourhood_grouped</i>	Barrio del alojamiento agrupado en categorías con suficiente número de observaciones.	Catagórica transformada en <i>dummy variables</i>	Permite capturar diferencias territoriales en el precio sin introducir categorías con escasa representatividad estadística.

Fuente: Elaboración propia.

Conviene destacar la diferencia entre las variables *beds* y *guests_included*. Aunque ambas se encuentran relacionadas con la capacidad del alojamiento, no representan exactamente la misma dimensión. Mientras que *beds* mide una característica física del inmueble, *guests_included* refleja una decisión comercial del anfitrión sobre el número de huéspedes incluidos en el precio base. Por tanto, la inclusión conjunta de ambas variables permite diferenciar entre capacidad física y política de *pricing* del alojamiento.

Junto a las variables seleccionadas, conviene justificar la exclusión del resto. De las 85 variables originales, la mayoría se descartó por motivos metodológicos: identificadores y metadatos de gestión sin relación económica con el precio; campos de texto libre no cuantificables sin procesamiento de lenguaje natural; variables redundantes con las ya incluidas, como el código postal, las coordenadas o los subcomponentes de la valoración, que habrían introducido multicolinealidad; y variables con demasiados valores faltantes (*accommodates*, *cancellation_policy*) o variabilidad casi nula (*bedrooms*). La siguiente tabla las resume junto con el motivo de su exclusión.

Tabla 2. Variables descartadas y motivo de exclusión

Categoría	Variables	Motivo de exclusión
Identificadores y metadatos	<i>id</i> , <i>host_id</i> , <i>host_name</i> , <i>host_since</i> , <i>host_location</i> , <i>host_response_rate</i> , etc.	Atributos de gestión sin relación económica directa con el precio del anuncio.
Texto libre	<i>name</i> , <i>summary</i> , <i>space</i> , <i>description</i> , <i>house_rules</i>	No cuantificables sin técnicas de procesamiento de lenguaje natural.

Categoría	Variables	Motivo de exclusión
Redundantes con las incluidas	zipcode, latitude, longitude, property_type y los subcomponentes de la valoración (review_scores_cleanliness, _location, etc.)	Miden información ya recogida por otras variables; el análisis de correlaciones confirma su escasa relación con el precio ($ r < 0,12$) y su redundancia, lo que generaría multicolinealidad.
Exceso de faltantes o datos imputados	accommodates (23,5 % NaN y 78,4 % de valores imputados), cancellation_policy (55,8 % NaN)	Datos ausentes o imputados que comprometen la fiabilidad de la estimación.
Variabilidad casi nula	bedrooms (96 % de los anuncios con valor 1)	Apenas varía entre anuncios, por lo que no explica diferencias de precio.

Fuente: Elaboración propia.

Esta selección se apoyó en un análisis de correlaciones entre las variables numéricas del conjunto original y el precio, cuyo código se recoge en el anexo. Dicho análisis respalda las exclusiones de la tabla anterior: las variables descartadas por redundancia presentan correlaciones con el precio muy débiles (en valor absoluto inferiores a 0,12) y/o una elevada correlación entre sí, mientras que las finalmente seleccionadas son las que muestran una relación más clara con el precio.

3.5. Especificación del modelo econométrico

Para analizar la relación entre las variables explicativas y el precio del alojamiento se utilizó un modelo de regresión lineal múltiple estimado mediante mínimos cuadrados ordinarios (OLS). La regresión lineal constituye una de las herramientas más utilizadas en econometría aplicada para estudiar relaciones entre variables económicas y cuantificar el efecto de distintos factores sobre una variable dependiente (Wooldridge, 2020).

La variable dependiente es *log_price*, es decir, $\log(1 + \text{precio})$. La elección de esta variable dependiente responde tanto a motivos estadísticos como económicos. Desde el punto de vista estadístico, permite reducir la asimetría del precio original. Desde el punto de vista económico, facilita la interpretación porcentual de los coeficientes.

La especificación general del modelo puede expresarse de la siguiente forma:

$$\begin{aligned}
\log(1 + price_i) = & \beta_0 + \beta_1 guests_included_i + \beta_2 beds_i + \beta_3 bathrooms_i + \beta_4 \\
& amenities_count_i \\
& + \beta_5 review_scores_rating_i + \beta_6 number_of_reviews_i + \sum_{k=1} \delta_k property_type_grouped_{ik} \\
& + \sum_{j=1} \gamma_j neighbourhood_grouped_{ij} + \varepsilon_i
\end{aligned}$$

donde $\log(1 + price_i)$ representa el logaritmo de uno más el precio del alojamiento i , las variables del lado derecho representan el conjunto de variables explicativas incluidas en el modelo, β recoge el efecto estimado de cada variable continua, γ_j los efectos asociados a las dummies de barrio, δ los efectos de las dummies de tipo de propiedad (apartamento como referencia) y ε_i representa el término de error.

La regresión lineal múltiple resulta adecuada para este trabajo porque permite obtener un modelo interpretable. A diferencia de otros modelos más complejos, como algunos algoritmos de aprendizaje automático, la regresión permite conocer el signo, la magnitud y la significatividad de cada variable. Esto resulta especialmente importante en un TFG cuyo objetivo no es únicamente predecir precios, sino comprender qué factores influyen en ellos.

Además, el modelo se estimó utilizando errores estándar robustos HC3. Esta decisión se debe a que en datos de mercado como los de Airbnb puede existir heterocedasticidad, es decir, que la variabilidad de los errores no sea constante para todos los niveles de precio. Los errores estándar robustos no modifican los coeficientes estimados, pero hacen que los contrastes de significatividad sean más fiables.

3.6. Evaluación predictiva

Aunque el objetivo principal del trabajo es explicativo, también se incorporó una evaluación de la capacidad predictiva del modelo. Esta decisión responde a la necesidad de comprobar si el modelo estimado es capaz de generalizar a observaciones no utilizadas durante el entrenamiento.

Para ello, la muestra final de 1.438 observaciones se dividió aleatoriamente en dos subconjuntos mediante la función *train_test_split* de *scikit-learn*, utilizando *random_state=42* para garantizar la reproducibilidad. El 80 % de las observaciones (1.150) se utilizó como conjunto de entrenamiento y el 20 % restante (288) como conjunto de prueba. El modelo se reestimó únicamente con los datos de entrenamiento y posteriormente se generaron predicciones sobre el conjunto de prueba.

Tabla 3. Partición de la muestra: entrenamiento y prueba

Conjunto	Observaciones	Porcentaje	Uso
Muestra total	1.438	100 %	Base de datos final tras limpieza
Conjunto de entrenamiento (train)	1.150	80 %	Estimación de los coeficientes del modelo
Conjunto de prueba (test)	288	20 %	Validación predictiva fuera de muestra

Fuente: Elaboración propia.

Esta metodología permite evaluar el comportamiento del modelo fuera de muestra. Es decir, no se analiza solo si el modelo ajusta bien los datos con los que fue estimado, sino también si es capaz de aproximar razonablemente el precio de alojamientos que no ha utilizado para aprender los coeficientes.

La capacidad predictiva se evaluó mediante dos métricas: RMSE y MAE. El RMSE mide la raíz del error cuadrático medio y penaliza más los errores grandes, mientras que el MAE mide el error absoluto medio y ofrece una interpretación más directa del error promedio.

Dado que la variable dependiente se encuentra expresada en logaritmos, estas métricas pueden interpretarse aproximadamente en términos porcentuales. Por tanto, la validación predictiva permite complementar el análisis explicativo con una aproximación a la utilidad práctica del modelo para estimar precios en nuevos casos.

No obstante, esta evaluación debe interpretarse con cautela. El modelo no pretende ser una herramienta automática de fijación exacta de precios, sino un instrumento analítico que permite identificar patrones relevantes y evaluar de forma aproximada su capacidad de predicción. Esta combinación de explicación e interpretación predictiva resulta coherente con el enfoque de *Business Analytics* adoptado en el trabajo.

4. Análisis e interpretación de resultados

Una vez estimado el modelo econométrico y completado el proceso de validación, este capítulo presenta el análisis e interpretación de los principales resultados obtenidos. El objetivo de esta sección no es únicamente describir los resultados estadísticos del modelo, sino interpretar qué implicaciones tienen desde una perspectiva económica y de negocio dentro del mercado de Airbnb en Madrid.

El análisis desarrollado combina distintas aproximaciones. En primer lugar, se realiza un análisis exploratorio de la variable precio y de algunas relaciones relevantes entre las variables incluidas en el modelo. Posteriormente, se interpretan los resultados de la regresión lineal múltiple, prestando especial atención tanto a la magnitud como a la significatividad estadística de los coeficientes estimados. Finalmente, se analizan distintos gráficos de diagnóstico y se evalúa la capacidad predictiva del modelo mediante validación fuera de muestra.

El propósito general es determinar hasta qué punto las variables seleccionadas permiten explicar el comportamiento del precio y si el modelo es capaz de capturar patrones relevantes del mercado.

4.1. Análisis exploratorio de la variable precio

Antes de interpretar los resultados del modelo, resulta necesario analizar el comportamiento de la variable dependiente utilizada en el estudio. En este caso, la variable de interés es el precio de los alojamientos publicados en Airbnb.

El análisis inicial del precio original muestra una distribución claramente asimétrica hacia la derecha. La mayoría de los alojamientos se concentra en niveles de precio relativamente bajos o medios, mientras que existe un grupo reducido de anuncios con precios considerablemente superiores al resto.

Este comportamiento es habitual en mercados turísticos y en plataformas digitales de alojamiento, donde la muestra utilizada se concentra principalmente en alojamientos económicos e intermedios, especialmente habitaciones privadas, aunque siguen existiendo diferencias relevantes de precio entre observaciones. La presencia de valores extremos implica que la distribución del precio no sigue una forma normal y presenta una elevada dispersión.

Desde un punto de vista estadístico, este tipo de distribución puede generar problemas en modelos lineales tradicionales. Los valores extremos pueden influir de forma excesiva sobre la estimación de los coeficientes y dificultar el cumplimiento de algunos supuestos del modelo, especialmente los relacionados con la homocedasticidad y la normalidad de los residuos.

Con el objetivo de reducir el efecto de esta asimetría y mejorar el comportamiento estadístico del modelo, se aplicó una transformación logarítmica sobre el precio mediante la creación de la variable *log_price*.

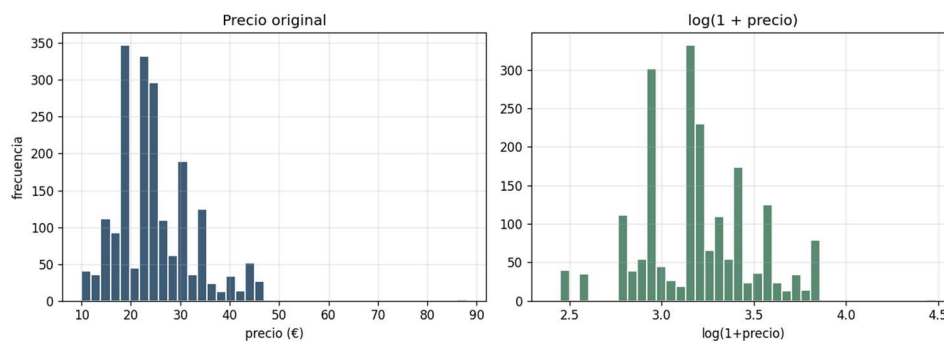
Tras aplicar esta transformación, la distribución presenta una forma más equilibrada y cercana a la normalidad. El uso del logaritmo permite reducir el impacto de los valores extremos y facilita la estimación de relaciones más estables entre las variables explicativas y la variable dependiente.

Para contrastar de forma objetiva la conveniencia de la transformación logarítmica, se aplicó el test de Shapiro-Wilk a ambas distribuciones, la del precio original y la de $\log(1 + \text{precio})$, complementándolo con los coeficientes de asimetría y curtosis. El precio original presenta una distribución marcadamente asimétrica y de colas pesadas ($W = 0,918$; asimetría = 1,25; curtosis = 3,93), mientras que la transformación logarítmica reduce de forma sustancial ambos rasgos ($W = 0,980$; asimetría = 0,13; curtosis = 0,10), aproximando la distribución a la normal.

En ambos casos el test rechaza formalmente la hipótesis de normalidad ($p < 0,001$). Este resultado es esperable dado el elevado tamaño muestral ($N = 1.993$), ya que el test de Shapiro-Wilk detecta como significativas desviaciones mínimas cuando la muestra es grande. Por ello, más que el p -valor, resultan informativos el estadístico W , que se acerca a 1 tras la transformación, y la fuerte reducción de la asimetría y la curtosis. En conjunto, la evidencia confirma que $\log(1 + \text{precio})$ es notablemente más simétrica y próxima a la normalidad que el precio original, lo que respalda su uso como variable dependiente del modelo.

La utilización de una especificación log-lineal resulta además coherente con gran parte de la literatura académica sobre precios hedónicos y análisis inmobiliario.

Figura 1. Distribución del precio original y del logaritmo del precio



Fuente: Elaboración propia a partir de Messaadi (2025).

4.2. Relación entre amenities y precio

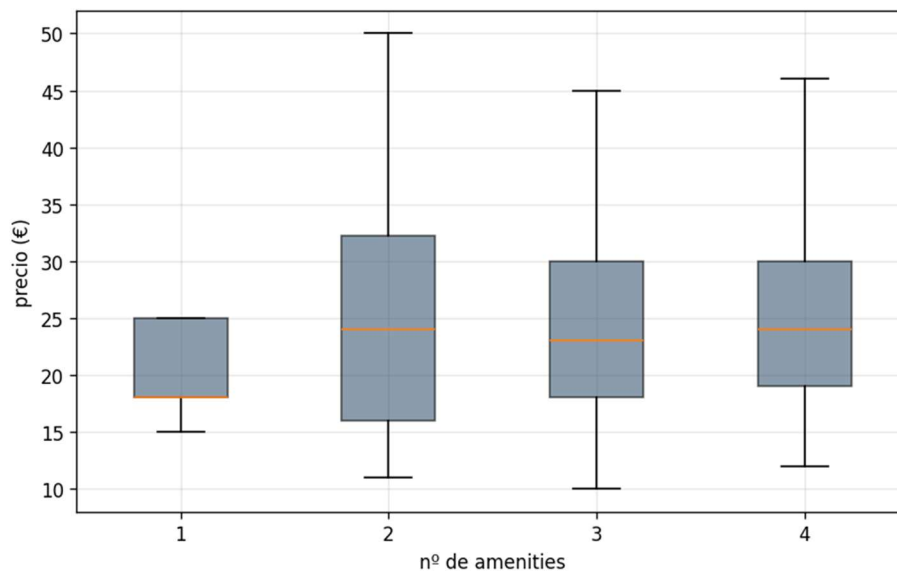
Una vez analizada la distribución de la variable dependiente, se estudió la relación entre algunas variables explicativas y el precio de los alojamientos.

Una de las relaciones exploradas fue la existente entre el número de servicios ofrecidos por el alojamiento (*amenities_count*) y el precio. La variable *amenities_count* se construyó como una aproximación al nivel de equipamiento del alojamiento, contabilizando el número total de servicios disponibles anunciados en la plataforma.

Inicialmente, se analizó la relación entre `amenities_count` y `log_price` mediante un gráfico de dispersión. Aunque este análisis permitía intuir cierta asociación, la elevada superposición de observaciones dificultaba interpretar con claridad cómo variaba el precio en función del número de amenities. Por este motivo, se complementó el análisis mediante un boxplot, que permite representar la distribución del precio para distintos niveles de equipamiento del alojamiento.

El boxplot revela que la relación entre el número de servicios y el precio es débil y no estrictamente creciente. La única diferencia apreciable se produce al pasar de uno a dos servicios: los anuncios con un solo servicio presentan una mediana de precio claramente inferior a la del resto. A partir de ahí, sin embargo, la relación se aplanan: las medianas de los grupos con dos, tres y cuatro servicios son prácticamente idénticas (en torno a 23-24 €) e incluso descienden ligeramente al pasar de dos a tres. Además, los rangos intercuartílicos de estos tres grupos se solapan casi por completo, lo que indica que sus distribuciones de precio son difícilmente distinguibles.

Figura 2. Distribución del precio según el número de amenities (boxplot)



Fuente: Elaboración propia.

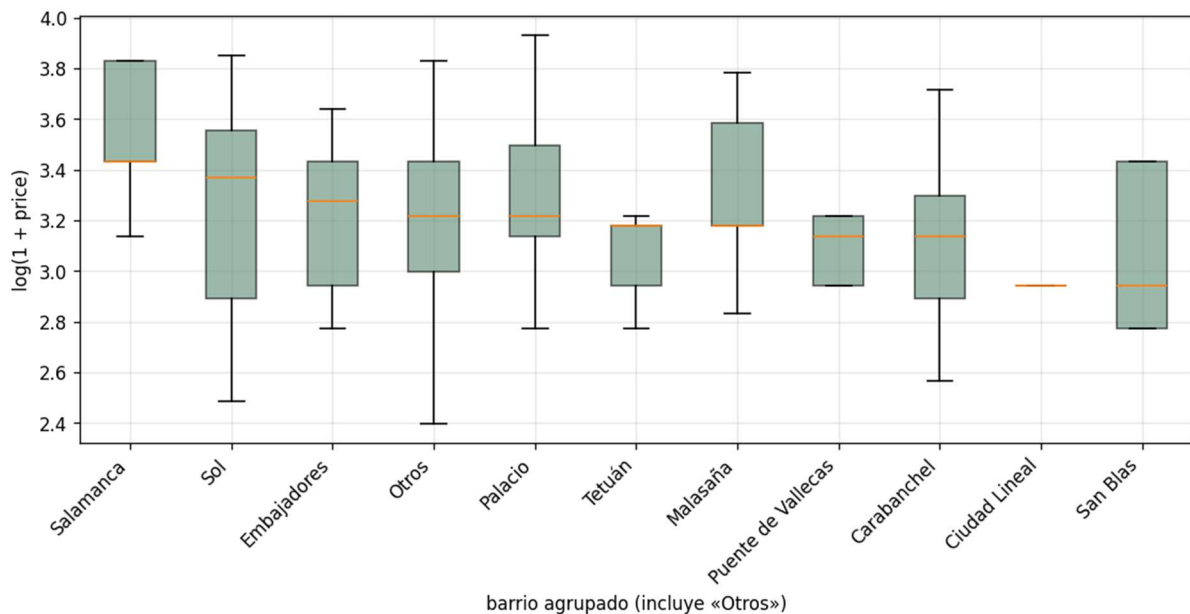
Por tanto, los datos no respaldan una relación positiva y creciente entre el equipamiento y el precio más allá del salto inicial de uno a dos servicios. Este patrón resulta coherente con el alcance de la variable comentado en el apartado anterior: al estar el campo limitado a un máximo de cuatro servicios y concentrarse la mayoría de los anuncios en los niveles superiores, `amenities_count` apenas permite discriminar diferencias de precio entre alojamientos bien equipados.

Desde el punto de vista económico, esta evidencia admite una lectura coherente: lo que penaliza el precio no es disponer de pocos extras, sino tratarse de un alojamiento muy básico, con un único servicio registrado. Una vez cubiertas las prestaciones mínimas que un huésped espera encontrar, añadir servicios adicionales deja de traducirse en un mayor precio en esta muestra. Este resultado matiza la hipótesis inicial y confirma la necesidad de interpretar la variable con cautela.

4.3. Diferencias territoriales en el precio

La localización constituye uno de los principales determinantes del precio tanto en mercados inmobiliarios tradicionales como en plataformas de alojamiento turístico. Por este motivo, se analizaron las diferencias territoriales en el precio entre los barrios más frecuentes de la muestra.

Figura 3. Distribución de $\log(1+price)$ por barrio agrupado



Fuente: Elaboración propia.

El gráfico muestra diferencias territoriales relevantes en la distribución del precio entre barrios de Madrid. En términos generales, algunas zonas presentan medianas de $\log(price)$ superiores a otras, lo que sugiere la existencia de heterogeneidad espacial dentro del mercado Airbnb madrileño.

En particular, barrios como Salamanca muestran niveles de precio relativamente más elevados, así como una mayor dispersión en determinadas observaciones. Este resultado resulta coherente con las características socioeconómicas de la zona, tradicionalmente asociada a niveles de renta superiores y a una mayor demanda de alojamiento. No obstante, también pueden observarse diferencias entre otros barrios incluidos en la muestra, reflejando que el precio no responde únicamente a características internas del alojamiento, sino también a factores territoriales.

Asimismo, el gráfico pone de manifiesto diferencias en la variabilidad del precio entre zonas. Mientras algunos barrios presentan distribuciones relativamente homogéneas, otros muestran una dispersión más amplia, lo que podría reflejar una mayor heterogeneidad en el tipo de alojamientos ofertados dentro de una misma área.

La categoría *Otros* agrupa barrios con menor representación individual dentro de la muestra y presenta una distribución intermedia del precio, lo que sugiere que los barrios menos frecuentes no siguen necesariamente un patrón homogéneo, sino que combinan características diversas.

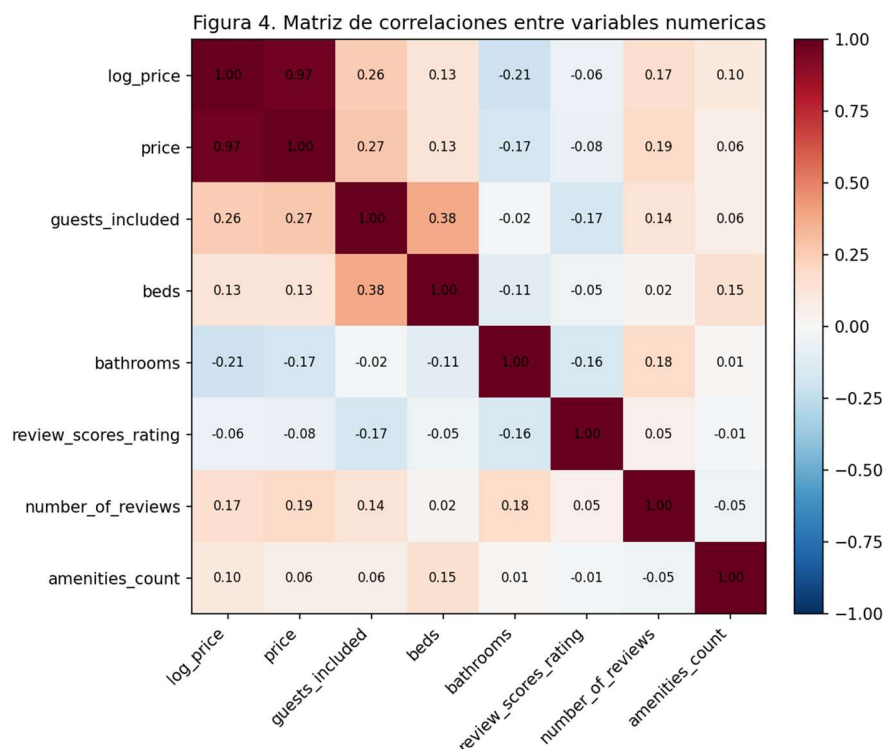
En conjunto, el análisis visual proporciona evidencia preliminar de que la localización desempeña un papel relevante en la fijación del precio de los alojamientos anunciados en Airbnb. Estas diferencias territoriales justifican la inclusión de variables de localización en el modelo econométrico mediante variables *dummy*, permitiendo controlar el efecto específico del barrio sobre el precio del alojamiento.

Además, estos resultados son coherentes con la literatura revisada anteriormente, donde múltiples estudios identifican la ubicación como uno de los principales determinantes del precio en Airbnb.

4.4. Análisis de correlaciones

Antes de estimar el modelo econométrico, se analizó la relación entre las principales variables numéricas incluidas en el estudio mediante una matriz de correlaciones. Este análisis permite identificar asociaciones lineales entre variables y detectar posibles problemas de multicolinealidad entre variables explicativas.

Figura 4. Matriz de correlaciones entre variables numéricas



Fuente: Elaboración propia.

La matriz de correlaciones muestra asociaciones generalmente moderadas entre las variables seleccionadas, sin evidencias de relaciones excesivamente elevadas que sugieran un problema grave de multicolinealidad. Este resultado resulta relevante, ya que una correlación excesivamente alta entre variables explicativas podría dificultar la estimación estable de los coeficientes del modelo.

En relación con la variable dependiente, se observa que algunas variables presentan asociaciones positivas con *log_price*. En particular, variables relacionadas con la capacidad del alojamiento, como *beds*, *bathrooms* y *guests_included*, muestran correlaciones positivas, lo que sugiere que alojamientos con mayor capacidad tienden a asociarse con niveles de precio superiores.

No obstante, aunque ambas variables se encuentran relacionadas con la capacidad del alojamiento, no representan exactamente la misma dimensión económica. Mientras que *beds* aproxima una característica física del inmueble, *guests_included* refleja una decisión comercial del anfitrión sobre el número de huéspedes incluidos en el precio base (puede que una habitación con dos camas la puedan reservar uno o dos huéspedes y a precios diferentes). Esta diferencia conceptual ayuda a explicar por qué ambas variables presentan una correlación positiva entre sí sin resultar completamente equivalentes.

Asimismo, la variable *amenities_count* presenta una asociación positiva con el precio, resultado coherente con la hipótesis de que un mayor nivel de equipamiento puede incrementar el valor percibido por el huésped.

En cambio, la variable *bathrooms* muestra una correlación negativa con el precio (en torno a 0,21), resultado coherente con el carácter compartido de las habitaciones analizadas: un mayor número de baños suele corresponder a viviendas más grandes y compartidas, en las que el precio de la habitación individual tiende a ser inferior.

Por otro lado, también se observan correlaciones moderadas entre algunas variables explicativas, especialmente entre aquellas relacionadas con la capacidad y el tamaño del alojamiento. Sin embargo, dichas relaciones no alcanzan niveles suficientemente elevados como para considerar la existencia de problemas graves de multicolinealidad.

En conjunto, el análisis de correlaciones proporciona evidencia preliminar sobre posibles asociaciones entre variables y permite justificar la selección de variables utilizada posteriormente en el modelo econométrico. No obstante, debe señalarse que la correlación no implica causalidad, por lo que resulta necesario recurrir al análisis multivariante para estimar el efecto individual de cada variable manteniendo constantes el resto de factores.

4.5. Resultados del modelo econométrico

Una vez completado el análisis exploratorio, se procedió a estimar un modelo de regresión lineal múltiple mediante mínimos cuadrados ordinarios (*Ordinary Least Squares, OLS*),

utilizando como variable dependiente el logaritmo del precio del alojamiento (*log_price*). Asimismo, se emplearon errores estándar robustos HC3 con el fin de corregir posibles problemas de heterocedasticidad y obtener inferencias estadísticas más fiables.

La Tabla 4 presenta los principales resultados del modelo, incluyendo coeficientes estimados, errores estándar robustos, *p*-valores y nivel de significatividad estadística.

Tabla 4. Resultados del modelo de regresión lineal múltiple (variable dependiente: log_price)

Variable	Coefficiente (β)	Error estándar (HC3)	p-valor	Significatividad
<i>guests_included</i>	0,1262	0,010	< 0,001	*** Significativa (positiva)
<i>beds</i>	-0,0041	0,019	0,832	No significativa
<i>bathrooms</i>	-0,1853	0,024	< 0,001	*** Significativa (negativa)
<i>amenities_count</i>	0,1071	0,008	< 0,001	*** Significativa (positiva)
<i>review_scores_rating</i>	-0,0723	0,014	< 0,001	*** Significativa (negativa)
<i>number_of_reviews</i>	0,0006	0,000	< 0,001	*** Significativa (positiva)
<i>Casa (tipo de propiedad)</i>	0,2933	0,030	< 0,001	*** Significativa (positiva)
<i>Otros tipos (hotel, hostal, etc.)</i>	0,3752	0,035	< 0,001	*** Significativa (positiva)
<i>Salamanca (barrio)</i>	0,4596	0,045	< 0,001	*** Significativa (positiva)
<i>Palacio</i>	0,0658	0,036	0,069	No significativa
<i>Tetuán</i>	0,0331	0,024	0,160	No significativa
<i>Embajadores</i>	-0,0037	0,030	0,902	No significativa
<i>Otros (resto barrios)</i>	-0,0155	0,022	0,481	No significativa
<i>San Blas</i>	-0,0249	0,035	0,475	No significativa
<i>Malasaña</i>	-0,0966	0,032	0,002	** Significativa (negativa)
<i>Sol</i>	-0,0997	0,039	0,010	** Significativa (negativa)
<i>Ciudad Lineal</i>	-0,1179	0,030	< 0,001	*** Significativa (negativa)
<i>Puente de Vallecas</i>	-0,3480	0,033	< 0,001	*** Significativa (negativa)

Notas: *** $p < 0,001$; ** $p < 0,01$. Errores estándar robustos HC3. Categorías de referencia, omitidas en la codificación dummy con *drop_first*: barrio = Carabanchel; tipo de propiedad = apartamento (Private room in apartment). Fuente: Elaboración propia.

El modelo presenta un coeficiente de determinación R^2 de 0,478 y un R^2 ajustado de 0,472, lo que indica que explica aproximadamente un 47,8 % de la variabilidad de $\log(\text{price})$.

Asimismo, el contraste global del modelo resulta estadísticamente significativo ($\text{Prob} > F < 0,001$), lo que indica que, consideradas conjuntamente, las variables explicativas incluidas presentan capacidad para explicar diferencias en el precio de los alojamientos.

En relación con las características del alojamiento, la variable *guests_included* presenta un coeficiente positivo y estadísticamente significativo ($\beta = 0,1262$; $p < 0,001$). Dado que la variable dependiente se encuentra expresada en logaritmos, este coeficiente puede interpretarse aproximadamente como un incremento del 13,5 % en el precio asociado a cada huésped adicional incluido en el precio base del alojamiento, manteniendo constantes el resto de factores. Este resultado sugiere que una mayor capacidad comercial del alojamiento se asocia con precios relativamente superiores.

Por el contrario, la variable *beds* no resulta estadísticamente significativa ($p = 0,832$). Una vez controlada la capacidad comercial mediante *guests_included*, el número de camas no aporta información adicional relevante sobre el precio.

En cuanto al nivel de equipamiento del alojamiento, la variable *amenities_count* presenta un efecto positivo y altamente significativo ($\beta = 0,1071$; $p < 0,001$). Esto implica que un incremento en el número de servicios ofrecidos se encuentra asociado con precios relativamente superiores, reforzando parcialmente la evidencia observada en el análisis exploratorio. Desde una perspectiva económica, este resultado resulta coherente con la idea de que un mayor nivel de equipamiento incrementa el valor percibido por el huésped.

La variable *bathrooms* presenta un coeficiente negativo y altamente significativo ($\beta = -0,1853$; $p < 0,001$). Este resultado, lejos de ser contraintuitivo, es coherente con la naturaleza de la muestra, compuesta íntegramente por habitaciones privadas: el número de baños no describe la unidad alquilada, sino el inmueble que la contiene, de modo que un mayor número de baños señala una vivienda más grande y compartida, en la que la habitación individual tiende a ofrecerse a un precio inferior.

Respecto a las variables reputacionales, *number_of_reviews* presenta un efecto positivo y estadísticamente significativo, aunque de magnitud muy reducida ($\beta = 0,0006$; $p < 0,001$). Por su parte, *review_scores_rating* muestra un coeficiente negativo y significativo ($\beta = -0,0723$; $p < 0,001$), lo que indica que una mayor valoración media no se traduce en un precio más elevado una vez controlado el resto de factores.

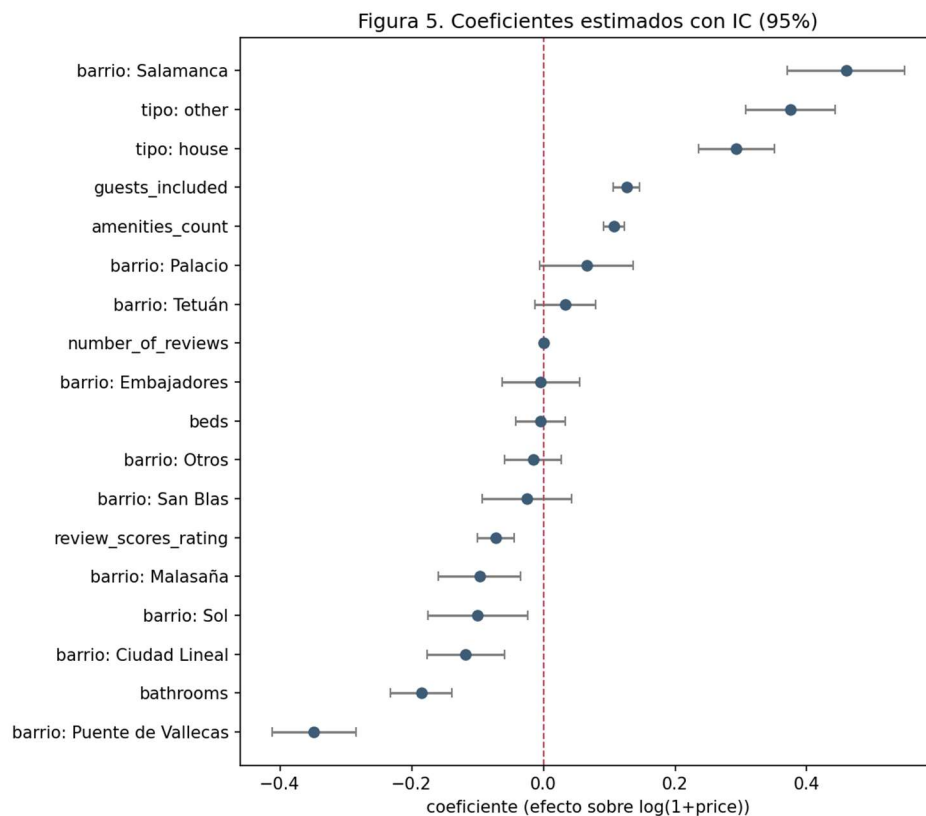
En relación con la tipología del alojamiento, el tipo de propiedad se incorpora como variable categórica agrupada en tres clases (apartamento, casa y otros tipos), tomando el apartamento como categoría de referencia. Las habitaciones situadas en casas presentan un precio significativamente superior al de los apartamentos ($\beta = 0,2933$; $p < 0,001$), equivalente a un incremento aproximado del 34 % una vez controladas las demás variables; la categoría «otros

tipos» (que agrupa hoteles, hostales, bed and breakfast y alojamientos similares) muestra un efecto todavía mayor ($\beta = 0,3752$; $p < 0,001$), en torno a un 46 % por encima del apartamento. Conviene matizar que esta diferencia emerge al mantener constante el número de baños: en términos brutos, apartamentos y casas presentan medianas de precio muy similares, de modo que el efecto positivo del tipo «casa» se revela una vez descontado el tamaño del inmueble.

Las variables territoriales reflejan diferencias relevantes en función de la localización. Tomando Carabanchel como categoría de referencia, el barrio de Salamanca ($\beta = 0,4596$; $p < 0,001$) presenta precios marcadamente superiores, coherente con su carácter de zona de alto poder adquisitivo. Palacio muestra también un coeficiente positivo, aunque solo marginalmente significativo ($\beta = 0,0658$; $p = 0,069$), por lo que su diferencia respecto a Carabanchel no puede afirmarse con rotundidad.

Por el contrario, barrios como Puente de Vallecas ($\beta = -0,3480$), Ciudad Lineal ($\beta = -0,1179$), Sol ($\beta = -0,0997$) y Malasaña ($\beta = -0,0966$) muestran precios significativamente inferiores, mientras que Palacio, Tetuán, Embajadores, San Blas y la categoría «Otros» no presentan diferencias estadísticamente significativas respecto a Carabanchel.

Figura 5. Gráfico de coeficientes estimados con intervalos de confianza al 95 %



Fuente: Elaboración propia.

En conjunto, los resultados sugieren que el precio de los alojamientos anunciados en Airbnb Madrid depende de una combinación de factores relacionados con capacidad comercial, nivel

de equipamiento, trayectoria del alojamiento dentro de la plataforma y localización territorial. No obstante, la magnitud de estos efectos debe interpretarse con cautela, dado que parte de la variabilidad observada permanece sin explicar y pueden existir factores no observados como decoración, diseño interior o preferencias específicas de los usuarios que también influyan sobre la fijación del precio.

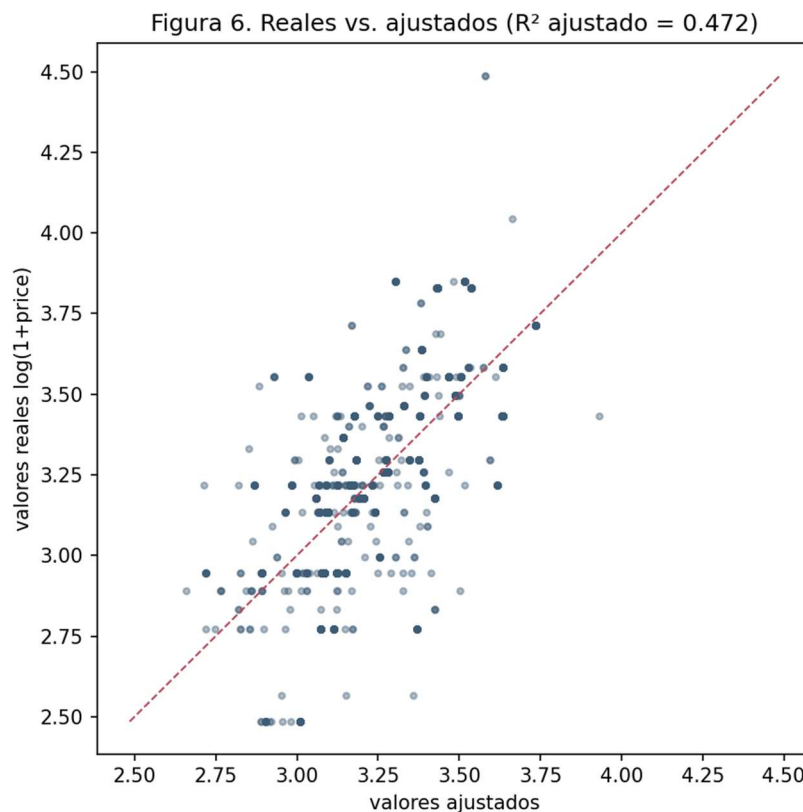
El gráfico de coeficientes permite visualizar claramente tanto la magnitud como la dirección de los efectos estimados, así como los intervalos de confianza asociados a cada variable. Esto facilita interpretar qué variables presentan efectos estadísticamente robustos y cuáles muestran una mayor incertidumbre.

4.6. Diagnóstico del modelo

Una vez estimado el modelo de regresión lineal múltiple, resulta necesario analizar si la especificación utilizada presenta un comportamiento razonablemente adecuado desde el punto de vista estadístico. Para ello, se realizaron distintos análisis de diagnóstico orientados a evaluar la calidad del ajuste y el comportamiento de los residuos del modelo.

En primer lugar, se compararon los valores reales observados con los valores ajustados por el modelo.

Figura 6. Valores reales vs. valores ajustados por el modelo



Fuente: Elaboración propia.

El gráfico muestra una relación positiva entre los valores reales observados y los valores ajustados por el modelo, lo que indica que este es capaz de capturar parcialmente la tendencia general presente en los datos. El ajuste presenta un coeficiente de determinación (R^2) de 0,478, indicando que aproximadamente un 47,8 % de la variabilidad observada en el logaritmo del precio puede explicarse mediante las variables incluidas en el modelo.

En este caso, aunque existe cierta dispersión, no se observan patrones claramente sistemáticos ni estructuras no lineales evidentes. Esto sugiere que la especificación lineal utilizada resulta razonable para aproximar la relación entre las variables incluidas y el precio.

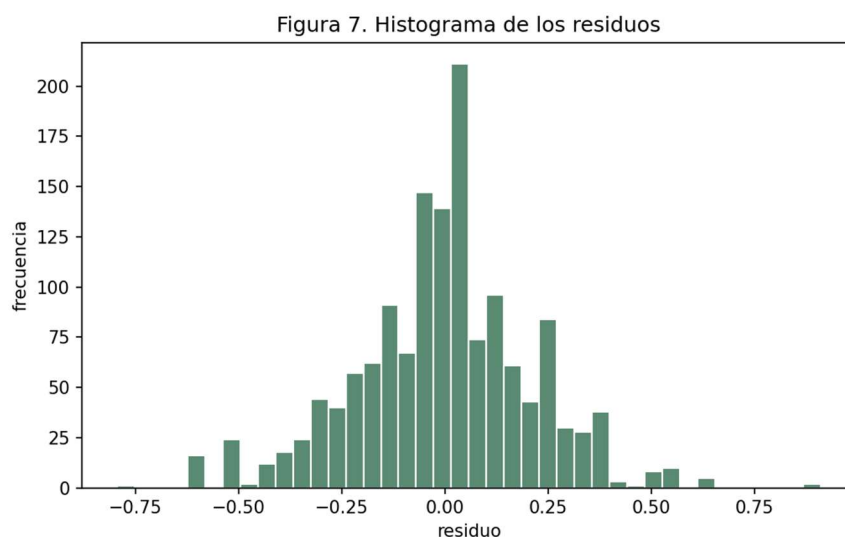
Por este motivo, el modelo fue estimado utilizando errores estándar robustos HC3, lo que permite obtener inferencias estadísticas más fiables incluso en presencia de desviaciones respecto al supuesto clásico de homocedasticidad.

Para respaldar formalmente estas decisiones se aplicaron dos contrastes complementarios. El test de Breusch-Pagan rechaza la hipótesis de homocedasticidad ($LM = 307,40$; $p < 0,001$), lo que confirma la presencia de heterocedasticidad y justifica el uso de errores estándar robustos HC3. Por su parte, los factores de inflación de la varianza (VIF) de las variables explicativas continuas, todos inferiores a 1,2, descartan la existencia de multicolinealidad relevante y validan su inclusión conjunta en el modelo.

Adicionalmente, como análisis de robustez, se reestimó el modelo desagregando el tipo de propiedad en sus once categorías originales mediante variables dummy, en lugar de la agrupación en tres clases empleada en el modelo principal. El efecto de *bathrooms* se mantuvo estable en signo y magnitud ($\beta = -0,201$) y la mejora del ajuste resultó marginal (el R^2 pasó de 0,478 a 0,503), lo que confirma la solidez del resultado y respalda la decisión de agrupar los tipos de propiedad para evitar categorías con muy pocas observaciones.

A continuación, se analizó la distribución de los residuos mediante un histograma.

Figura 7. Histograma de los residuos del modelo

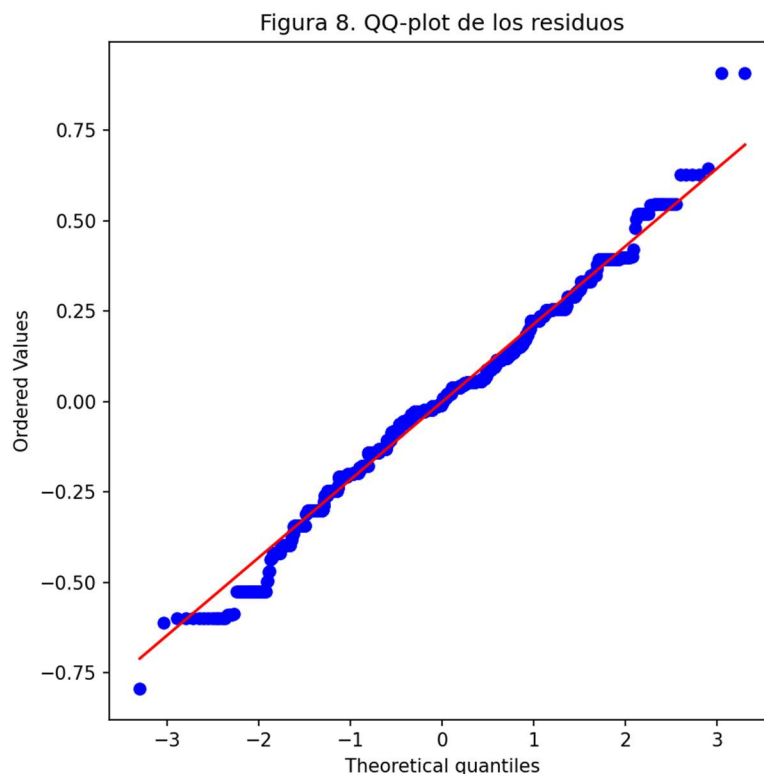


Fuente: Elaboración propia.

El histograma muestra que los residuos se encuentran relativamente centrados alrededor de cero, lo que indica que el modelo no presenta sesgos sistemáticos especialmente importantes en las predicciones.

Para complementar este análisis, se utilizó un gráfico QQ-plot que compara la distribución observada de los residuos con una distribución normal teórica.

Figura 8. QQ-plot de los residuos



Fuente: Elaboración propia.

En el QQ-plot puede observarse que gran parte de las observaciones se sitúa relativamente cerca de la línea diagonal, especialmente en la parte central de la distribución. Esto indica que los residuos presentan un comportamiento razonablemente próximo a la normalidad en los valores intermedios.

No obstante, también se observan desviaciones en los extremos de la distribución, particularmente en las colas. Estas desviaciones reflejan la existencia de algunas observaciones con errores de predicción relativamente elevados.

En conjunto, el análisis de diagnóstico sugiere que el modelo presenta un comportamiento razonablemente adecuado para un estudio aplicado de *Business Analytics*. Aunque existen

limitaciones y parte de la variabilidad del precio permanece sin explicar, el modelo consigue capturar relaciones económicas relevantes y ofrece resultados interpretables y consistentes.

4.7. Evaluación predictiva del modelo

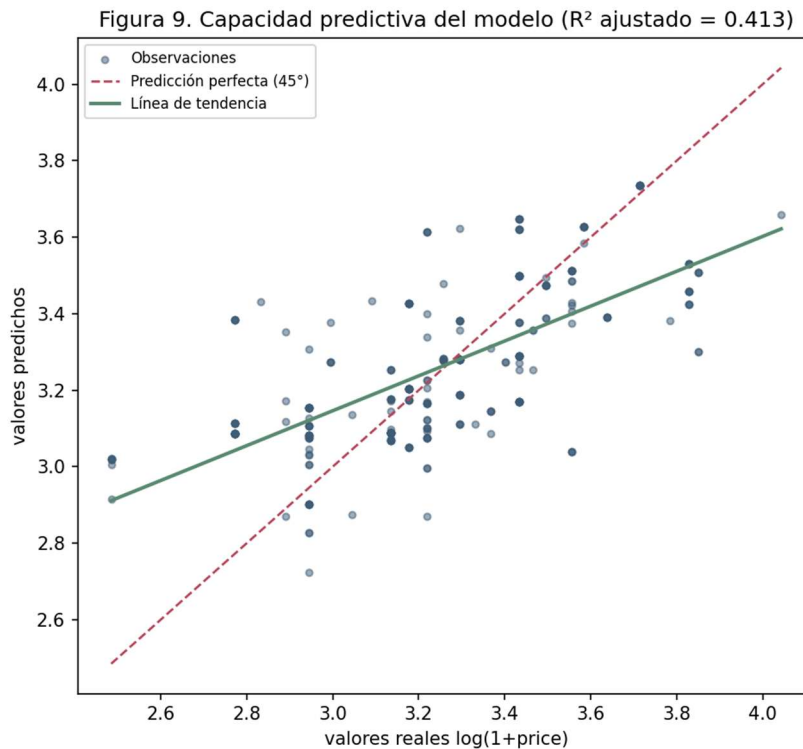
Además del análisis explicativo desarrollado mediante la regresión lineal múltiple, se llevó a cabo una evaluación de la capacidad predictiva del modelo. La incorporación de esta fase responde a la necesidad de comprobar hasta qué punto el modelo es capaz de generalizar sobre observaciones no utilizadas durante el proceso de estimación.

Este aspecto resulta especialmente relevante en el contexto de *Business Analytics*, donde no solo interesa interpretar relaciones entre variables, sino también evaluar si el modelo puede aproximar razonablemente el comportamiento de nuevos datos.

La utilización de conjuntos de entrenamiento y prueba constituye una práctica habitual en modelos predictivos y aprendizaje estadístico, ya que permite evaluar la capacidad de generalización del modelo sobre datos no utilizados durante el proceso de estimación (James et al., 2021).

Con el objetivo de evaluar la capacidad predictiva del modelo fuera de la muestra utilizada para su estimación, se implementó un procedimiento de validación mediante partición de datos (*train-test split*). Para ello, la base de datos final, compuesta por 1.438 observaciones, se dividió aleatoriamente en un conjunto de entrenamiento (1.150 observaciones; 80 %) y un conjunto de prueba (288 observaciones; 20 %).

El modelo fue reestimado utilizando únicamente el conjunto de entrenamiento, mientras que el conjunto de prueba se reservó para evaluar la capacidad de predicción sobre observaciones no utilizadas durante la estimación.

Figura 9. Valores reales vs. valores predichos sobre el conjunto de prueba

Fuente: Elaboración propia.

El gráfico muestra una relación positiva entre los valores reales y los valores predichos, lo que indica que el modelo es capaz de capturar parcialmente la estructura general presente en los datos. El coeficiente de determinación obtenido sobre el conjunto de prueba ($R^2 = 0,450$) indica que aproximadamente un 45,0 % de la variabilidad observada en el logaritmo del precio puede explicarse fuera de muestra, resultado relativamente próximo al R^2 obtenido durante la estimación sobre el conjunto de entrenamiento ($R^2 = 0,484$). La Figura 9 recoge además el R^2 ajustado sobre el conjunto de prueba (0,413), que corrige ese valor penalizando por el número de predictores incluidos en el modelo.

Esta diferencia relativamente reducida entre ambos valores sugiere que el modelo no presenta un problema severo de sobreajuste (*overfitting*), dado que mantiene una capacidad predictiva razonablemente estable al aplicarse sobre datos no utilizados previamente en la estimación.

No obstante, el gráfico también muestra un patrón característico en el comportamiento predictivo del modelo. En términos generales, se observa cierta tendencia a sobreestimar algunos alojamientos con precios relativamente bajos y subestimar determinados alojamientos con precios más elevados, reflejando una capacidad limitada para capturar completamente los valores extremos de la distribución.

Adicionalmente, el rendimiento predictivo fue evaluado mediante dos métricas complementarias: el error absoluto medio (*Mean Absolute Error, MAE*) y la raíz del error

cuadrático medio (*Root Mean Squared Error, RMSE*). Los resultados obtenidos fueron de $MAE = 0,168$ y $RMSE = 0,223$.

Tabla 5. Métricas de ajuste y evaluación predictiva del modelo

Métrica	Valor	Interpretación
R ² (muestra completa)	0,478	47,8 % de la variabilidad de log(price) explicada.
R ² ajustado	0,472	Penaliza la inclusión de variables adicionales.
R ² (entrenamiento)	0,484	Modelo reestimado sobre 1.150 observaciones.
R ² (prueba)	0,450	Capacidad explicativa fuera de muestra; sin sobreajuste severo.
MAE (prueba)	0,168	Error absoluto medio sobre log(1 + price).
RMSE (prueba)	0,223	Raíz del error cuadrático medio; penaliza errores grandes.
Prob > F	< 0,001	Modelo globalmente significativo.

Fuente: Elaboración propia.

Dado que la variable dependiente se encuentra expresada en logaritmos, estas métricas deben interpretarse como errores medios sobre el logaritmo del precio y no directamente en euros. No obstante, los resultados sugieren que el modelo presenta un nivel de precisión razonable para un estudio aplicado sobre un mercado complejo y heterogéneo como Airbnb.

En conjunto, la validación predictiva indica que el modelo presenta una capacidad moderada para anticipar diferencias de precio entre alojamientos, manteniendo un rendimiento relativamente estable fuera de muestra. Aunque parte de la variabilidad permanece sin explicar, el modelo resulta suficientemente robusto para cumplir el objetivo analítico planteado en este trabajo.

5. Conclusiones

El presente capítulo recoge las principales conclusiones derivadas del análisis desarrollado a lo largo del trabajo. En primer lugar, se sintetizan los principales hallazgos obtenidos a partir del modelo econométrico y del análisis exploratorio realizado sobre los datos de Airbnb en Madrid. Posteriormente, se analizan las implicaciones prácticas y la utilidad del modelo desde una perspectiva de *Business Analytics*, así como las principales limitaciones del estudio y posibles líneas futuras de investigación.

El objetivo final de este capítulo es evaluar hasta qué punto el modelo desarrollado permite comprender el comportamiento del precio dentro del segmento de habitaciones privadas y alojamientos económicos de Airbnb Madrid, así como valorar la utilidad de las técnicas cuantitativas aplicadas en el contexto de plataformas digitales y mercados turísticos.

5.1. Principales hallazgos del estudio

Este Trabajo Fin de Grado ha analizado los principales determinantes del precio de los alojamientos anunciados en Airbnb Madrid mediante técnicas de Business Analytics, estimando un modelo de regresión lineal múltiple sobre una muestra final de 1.438 observaciones del segmento de habitaciones privadas y alojamientos económicos. Los resultados muestran que solo determinadas variables presentan una relación estadísticamente significativa con el precio una vez consideradas conjuntamente.

Entre las características del alojamiento, *guests_included* y *amenities_count* tienen un efecto positivo y significativo: más huéspedes incluidos en el precio base y un mayor equipamiento se asocian a precios más altos. Por el contrario, *bathrooms* presenta un efecto negativo y significativo, coherente con una muestra de habitaciones privadas, donde un mayor número de baños señala un inmueble más grande y compartido en el que la habitación individual se ofrece a un precio inferior. La variable *number_of_reviews* resulta positiva y significativa, lo que sugiere que los anuncios con mayor trayectoria y visibilidad tienden a precios algo superiores, mientras que *review_scores_rating* muestra un coeficiente negativo: una valoración media más alta no se traduce en un precio mayor una vez controlado el resto de variables.

El tipo de alojamiento, incorporado como variable categórica agrupada (apartamento, casa y otros tipos), también resulta relevante: las habitaciones en casas y, sobre todo, en otros tipos (hoteles, hostales y similares) son significativamente más caras que las situadas en apartamentos, una vez controlados el tamaño y el equipamiento.

Desde una perspectiva territorial, se observan diferencias significativas entre barrios: Salamanca presenta precios superiores a la categoría de referencia, mientras que Puente de Vallecas, Ciudad Lineal, Sol y Malasaña los presentan inferiores; el resto de barrios no muestra diferencias significativas. Asimismo, el uso del logaritmo del precio (*log_price*) mejoró el comportamiento estadístico de la estimación.

Finalmente, desde la perspectiva predictiva, el modelo mostró una capacidad moderada (R^2 fuera de muestra = 0,450; RMSE = 0,223; MAE = 0,168), sin evidencia de sobreajuste severo. En conjunto, el precio de los alojamientos analizados depende principalmente del número de huéspedes incluidos, el nivel de equipamiento, determinadas características del anuncio y la localización, si bien una parte relevante de su variabilidad permanece sin explicar, lo que apunta a factores no recogidos en la base de datos.

5.2. Implicaciones prácticas y utilidad del modelo

Más allá de los resultados puramente estadísticos, el modelo desarrollado presenta una utilidad relevante desde una perspectiva aplicada y de negocio. El análisis realizado permite identificar qué características se encuentran asociadas con precios más elevados dentro del segmento de habitaciones privadas y alojamientos económicos de Airbnb Madrid.

En particular, variables relacionadas con la capacidad del alojamiento, el número de servicios ofrecidos y la localización muestran una influencia significativa sobre el precio. Esto implica que los anfitriones pueden utilizar este tipo de análisis para comprender mejor qué elementos generan un mayor valor percibido por parte de los usuarios y cómo determinadas decisiones relacionadas con el equipamiento o la gestión del alojamiento pueden afectar al posicionamiento del anuncio dentro de la plataforma.

Asimismo, el trabajo pone de manifiesto la utilidad de las técnicas de *Business Analytics* como herramienta de apoyo para la toma de decisiones en mercados digitales. El uso de modelos cuantitativos permite transformar grandes volúmenes de datos en información interpretable y útil tanto para anfitriones individuales como para empresas vinculadas al sector turístico y plataformas digitales.

Aunque el modelo desarrollado no tiene como objetivo sustituir sistemas profesionales de *pricing* dinámico, sí constituye una aproximación útil para comprender las principales dinámicas económicas que intervienen en la formación de precios dentro de Airbnb.

Además, la incorporación de una fase de validación predictiva aporta un valor añadido importante al trabajo, ya que permite evaluar si las relaciones identificadas mantienen capacidad de generalización sobre nuevos datos y no únicamente sobre la muestra utilizada durante la estimación.

5.3. Limitaciones del trabajo

A pesar de los resultados obtenidos, el trabajo presenta diversas limitaciones que deben tenerse en cuenta al interpretar las conclusiones del estudio.

En primer lugar, la muestra utilizada se concentra principalmente en habitaciones privadas y alojamientos de rango económico e intermedio dentro de Airbnb Madrid. Por tanto, los resultados obtenidos no pueden extrapolarse directamente al conjunto completo del mercado

Airbnb ni a otros segmentos como apartamentos turísticos premium o viviendas completas de lujo.

Asimismo, el tamaño final de la muestra, aunque suficiente para estimar el modelo planteado, sigue siendo relativamente reducido en comparación con la dimensión real del mercado Airbnb. Esto limita parcialmente la capacidad de generalización de los resultados.

Otra limitación importante se relaciona con la disponibilidad de variables. Existen numerosos factores potencialmente relevantes para explicar el precio que no están recogidos en la base de datos utilizada. Entre ellos destacan elementos como la calidad real del inmueble, el diseño interior, la experiencia previa del anfitrión, la estacionalidad, la flexibilidad de cancelación o variables relacionadas con la demanda turística en momentos concretos.

Del mismo modo, el modelo empleado corresponde a una regresión lineal múltiple relativamente sencilla y orientada a la interpretabilidad de los resultados. Aunque esta aproximación resulta adecuada para los objetivos académicos del trabajo, existen técnicas más avanzadas de *machine learning* que podrían mejorar la capacidad predictiva del modelo.

No obstante, muchas de estas técnicas presentan una menor interpretabilidad económica, lo que dificulta comprender el efecto individual de cada variable sobre el precio.

Por último, debe señalarse que el mercado Airbnb presenta un comportamiento altamente dinámico, donde los precios pueden variar rápidamente en función de factores externos, eventos concretos o cambios en la demanda turística. Como consecuencia, los resultados obtenidos reflejan el comportamiento del mercado en el contexto temporal específico de la muestra analizada.

5.4. Líneas futuras de investigación

A partir de las limitaciones identificadas, pueden plantearse diversas líneas futuras de investigación que permitan ampliar o mejorar el análisis desarrollado en este trabajo.

En primer lugar, resultaría interesante ampliar la muestra utilizada incorporando otros tipos de alojamiento presentes en Airbnb, especialmente apartamentos completos y alojamientos de gama alta. Esto permitiría comparar si los determinantes del precio difieren entre distintos segmentos del mercado.

Asimismo, futuras investigaciones podrían incorporar nuevas variables explicativas relacionadas con aspectos cualitativos del alojamiento, reputación del anfitrión, estacionalidad o comportamiento de la demanda turística.

Otra posible línea de investigación consistiría en aplicar técnicas más avanzadas de *machine learning*, como Random Forest, XGBoost o redes neuronales, con el objetivo de comparar su capacidad predictiva respecto al modelo lineal utilizado en este trabajo.

Del mismo modo, podría analizarse la evolución temporal de los precios mediante modelos dinámicos o estudios longitudinales que permitan observar cómo cambian las estrategias de *pricing* en Airbnb a lo largo del tiempo.

Finalmente, también resultaría interesante ampliar el análisis a otras ciudades españolas o europeas para estudiar hasta qué punto los determinantes del precio presentan patrones similares o diferencias territoriales relevantes.

En conjunto, los resultados obtenidos permiten concluir que el precio de los alojamientos en Airbnb depende de una combinación de factores estructurales, reputacionales y territoriales, y que las técnicas de *Business Analytics* constituyen una herramienta útil para comprender este tipo de mercados digitales. El trabajo aporta así una aproximación aplicada al análisis del *pricing* en plataformas digitales y contribuye a una mejor comprensión del comportamiento del segmento de habitaciones privadas dentro de Airbnb Madrid.

6. Declaración de uso de herramientas de IAG

ADVERTENCIA: Desde la Universidad consideramos que ChatGPT u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo Fin de Grado para generar código, porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, Francisco Javier Meliá García, estudiante de *ADE y Business Analytics* de la Universidad Pontificia Comillas, al presentar mi Trabajo Fin de Grado titulado "*Análisis de los determinantes del precio en el segmento de habitaciones privadas de Airbnb Madrid: una aproximación desde Business Analytics*", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG sólo en el contexto de las actividades descritas a continuación:

- **Brainstorming de ideas de investigación:** se utilizó como apoyo para ordenar las primeras ideas del trabajo, explorar posibles enfoques y concretar mejor el tema de investigación.
- **Referencias:** se utilizó como ayuda para orientar la búsqueda bibliográfica y entender mejor cómo algunas fuentes podían relacionarse con el tema. La selección final de referencias fue realizada por la autora.
- **Metodología:** se utilizó como apoyo para aclarar la estructura metodológica y explicar mejor algunas decisiones tomadas durante el análisis.
- **Interpretación de código:** se utilizó como ayuda para revisar y comprender partes del código, así como para explicar de forma más clara los pasos técnicos realizados.
- **Corrección de estilo y lenguaje:** se utilizó para mejorar la redacción, la claridad y la coherencia del texto, manteniendo el contenido y las ideas principales del trabajo.
- **Revisión:** se utilizó como apoyo para revisar la estructura general del documento, detectar posibles repeticiones y mejorar la conexión entre secciones.
- **Apoyo en la interpretación de resultados:** se utilizó para expresar los resultados de forma más clara y comprensible, especialmente en las partes más técnicas del análisis.
- **Apoyo en la visualización y presentación de datos:** se utilizó como ayuda para pensar cómo presentar la información de manera más ordenada mediante tablas, gráficos o explicaciones más claras.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para qué se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 26/04/2026

Firma: _____ Francisco Javier Meliá García _____

7. Referencias

- Carrero Bosch, I., Fabra Florit, M. E. y Valor Martínez, C. (2012). *El libro de cabecera del investigador en Economía y Empresa. Una guía interactiva para el investigador novel y el investigador profesional*. Madrid: Universidad Pontificia Comillas.
- Dogru, T., Mody, M. y Suess, C. (2019). 'Adding evidence to the debate: Quantifying Airbnb's disruptive impact on ten key hotel markets', *Tourism Management*, 72, pp. 27–38.
- Ert, E., Fleischer, A. y Magen, N. (2016). 'Trust and reputation in the sharing economy', *Tourism Management*, 55, pp. 62–73.
- García, J. I., Garrido, R. y Hernández, N. (2003). *Libro de estilo*. 2ª ed. Madrid: Universidad Pontificia Comillas.
- Gibbs, C., Guttentag, D., Gretzel, U., Morton, J. y Goodwill, A. (2018). 'Pricing in the sharing economy: A hedonic pricing model applied to Airbnb listings', *Journal of Travel & Tourism Marketing*, 35(1), pp. 46–56.
- Guttentag, D. (2015). 'Airbnb: disruptive innovation and the rise of an informal tourism accommodation sector', *Current Issues in Tourism*, 18(12), pp. 1192–1217.
- James, G., Witten, D., Hastie, T. y Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R*. 2nd edn. New York: Springer.
- Messaadi, A. (2025). *Airbnb Private Rooms in Madrid, Spain*. Kaggle. Disponible en: <https://www.kaggle.com/datasets/abderrahmanemessaadi/airbnb-private-rooms-in-madrid-spain> (Consultado: 7 de mayo de 2026).
- Quattrone, G., Proserpio, D., Quercia, D., Capra, L. y Musolesi, M. (2016). 'Who benefits from the sharing economy of Airbnb?', *Proceedings of the 25th International World Wide Web Conference*, pp. 1385–1394.
- Rosen, S. (1974). 'Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition', *Journal of Political Economy*, 82(1), pp. 34–55.
- Wang, D. y Nicolau, J. L. (2017). 'Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb.com', *International Journal of Hospitality Management*, 62, pp. 120–131.
- Wooldridge, J. M. (2020). *Introductory Econometrics: A Modern Approach*. 7th edn. Boston: Cengage Learning.
- Zervas, G., Proserpio, D. y Byers, J. W. (2017). 'The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry', *Journal of Marketing Research*, 54(5), pp. 687–705.

Anexo. Código del análisis

Este anexo recoge el código completo del notebook de Python utilizado para realizar la limpieza de datos, la construcción de variables, la estimación del modelo de regresión lineal múltiple y la validación predictiva del trabajo. El código se ha desarrollado en Jupyter / Google Colab utilizando las librerías pandas, numpy, scipy, statsmodels, scikit-learn y matplotlib. Se reproduce a continuación, celda a celda, tal como figura en el notebook ejecutado sobre la base de datos madrid_.csv (Messaadi, 2025).

La estructura del notebook sigue una secuencia lógica que va desde la carga del archivo original hasta la guardado de los resultados finales del modelo. Cada paso se documenta con una breve explicación previa al bloque de código correspondiente.

TFG Determinantes del precio en Airbnb Madrid

Modelo hedónico (OLS, HC3) versión final con bathrooms

Notebook completo y reproducible: limpieza, construcción de variables, exportación del dataset limpio a Excel, análisis exploratorio (Shapiro-Wilk), las 9 figuras del TFG, modelo de regresión con errores robustos HC3, diagnóstico (con R^2 ajustado) y evaluación predictiva (con R^2 ajustado).

Bloque 1

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
from scipy import stats
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
import matplotlib.pyplot as plt

pd.set_option("display.max_columns", None)
plt.rcParams["figure.dpi"] = 110
```

1. Carga de datos

Bloque 2

```
df = pd.read_csv("madrid_.csv")
print("Dimensiones:", df.shape)
df.head()
```

1.1 Correlación de las variables numéricas del dataset original

Antes de la selección de variables, se examina la correlación entre las variables numéricas disponibles en el dataset original y el precio. Sirve para justificar qué variables se incorporan al modelo (las que presentan una relación más relevante con el precio y no son redundantes entre sí) y cuáles se descartan.

Bloque 3

```
# Correlación de las variables numéricas candidatas del dataset original
corr_vars = ["price", "guests_included", "accommodates", "beds", "bedrooms",
             "bathrooms", "number_of_reviews", "review_scores_rating",
```

```

        "review_scores_accuracy", "review_scores_cleanliness",
        "review_scores_checkin", "review_scores_communication",
        "review_scores_location", "review_scores_value"]

tmp = df[corr_vars].copy()
tmp["price"] = pd.to_numeric(tmp["price"].astype(str).str.replace("$", "",
regex=False)
                                .str.replace(",", "", regex=False).str.strip(),
errors="coerce")
tmp["guests_included"] = pd.to_numeric(tmp["guests_included"].astype(str)
                                .str.extract(r"(\d+)", expand=False),
errors="coerce")
for c in corr_vars:
    tmp[c] = pd.to_numeric(tmp[c], errors="coerce")
corr_all = tmp.corr(numeric_only=True)

fig, ax = plt.subplots(figsize=(10, 8)); im = ax.imshow(corr_all, cmap="RdBu_r",
vmin=-1, vmax=1)
ax.set_xticks(range(len(corr_vars))); ax.set_xticklabels(corr_vars, rotation=45,
ha="right")
ax.set_yticks(range(len(corr_vars))); ax.set_yticklabels(corr_vars)
for i in range(len(corr_vars)):
    for j in range(len(corr_vars)):
        ax.text(j, i, f"{corr_all.iloc[i, j]:.2f}", ha="center", va="center",
        fontsize=7)
plt.colorbar(im, fraction=0.046, pad=0.04); ax.grid(False)
plt.title("Correlacion entre las variables numericas del dataset original");
plt.tight_layout(); plt.show()

print("Correlacion de cada variable con el precio (ordenada por magnitud):")
print(corr_all["price"].drop("price").reindex(

corr_all["price"].drop("price").abs().sort_values(ascending=False).index).round(3).to
_string())

```

2. Limpieza y conversión de variables numéricas

Bloque 4

```

df["price"] = (df["price"].astype(str)
                .str.replace("$", "", regex=False)
                .str.replace(",", "", regex=False).str.strip())
df["price"] = pd.to_numeric(df["price"], errors="coerce")

df["guests_included"] = pd.to_numeric(
    df["guests_included"].astype(str).str.extract(r"(\d+)", expand=False),
    errors="coerce")

for c in ["beds", "review_scores_rating", "number_of_reviews", "bathrooms"]:
    df[c] = pd.to_numeric(df[c], errors="coerce")

# Eliminar filas duplicadas exactas
df = df.drop_duplicates().copy()

```

3. Construcción de variables analíticas

log_price, amenities_count, property_type_grouped (apartamento / casa / otro) y la agrupación de barrios.

Bloque 5

```
df["log_price"] = np.log1p(df["price"])

raw = df["amenities"].fillna("").astype(str)
df["amenities_count"] = raw.str.count(",") + 1
df.loc[df["amenities"].isna(), "amenities_count"] = 0

# Tipo de propiedad agrupado en 3 categorías: apartamento / casa / otro.
# Se evita codificar una variable nominal como continua (propertytypeid) y se
# agrupan los tipos con muy pocos casos, igual que con los barrios.
df["property_type_grouped"] = np.where(
    df["property_type"].eq("Private room in apartment"), "apartment",
    np.where(df["property_type"].eq("Private room in house"), "house", "other"))

top = df["neighbourhood_cleansed"].value_counts().head(10).index
df["neighbourhood_grouped"] = df["neighbourhood_cleansed"].where(
    df["neighbourhood_cleansed"].isin(top), "Otros")

df[["price", "log_price", "guests_included", "beds", "bathrooms",
    "amenities_count"]].describe()
```

3.1 Exportar el dataset limpio a Excel

Se guarda el conjunto depurado (variables analíticas) en un archivo .xlsx para su descarga. Requiere el paquete openpyxl (pip install openpyxl).

Bloque 6

```
cols_export = ["log_price", "price", "guests_included", "beds", "bathrooms",
               "review_scores_rating", "number_of_reviews", "property_type",
               "property_type_grouped", "amenities_count",
               "neighbourhood_cleansed", "neighbourhood_grouped"]
df_limpio = df[cols_export].copy()

ruta = "madrid_limpio.xlsx"
df_limpio.to_excel(ruta, index=False)
print(f"Dataset limpio exportado a '{ruta}' -> {df_limpio.shape[0]} filas,
{df_limpio.shape[1]} columnas")

# Descarga:
# - En Google Colab:      from google.colab import files; files.download(ruta)
# - En Jupyter local:    el archivo queda en la carpeta del notebook, listo para
                          abrir/descargar.
```

4. Valores faltantes de las variables del modelo

Bloque 7

```
vars_check = ["log_price", "guests_included", "beds", "bathrooms",
              "review_scores_rating", "number_of_reviews",
              "property_type_grouped", "amenities_count", "neighbourhood_cleansed"]
(df[vars_check].isna().mean() * 100).round(1).sort_values(ascending=False).rename("%
NaN")
```

5. Análisis exploratorio del precio

5.1 Test de Shapiro-Wilk: precio vs. log(1 + precio)

Bloque 8

```

precio = df["price"].dropna()
log_precio = np.log1p(precio)
print(f"{'distribucion':16s}{'N':>6}{'W':>10}{'p-valor':>13}{'asimetria':>12}{'kurtosis':>11}")
for nombre, serie in [("precio", precio), ("log(1+precio)", log_precio)]:
    W, p = stats.shapiro(serie)
    print(f"{'nombre':16s}{'len(serie)':>6}{'W':>10.4f}{'p':>13.2e}"
          f"{'stats.skew(serie)':>+12.3f}{'stats.kurtosis(serie)':>+11.3f}")

```

Figura 1. Distribución del precio original y del logaritmo del precio

Bloque 9

```

fig, ax = plt.subplots(1, 2, figsize=(11, 4))
ax[0].hist(precio, bins=40, color="#3E5C76", edgecolor="white");
ax[0].set_title("Precio original"); ax[0].set_xlabel("precio (€)")
ax[1].hist(log_precio, bins=40, color="#5B8A72", edgecolor="white");
ax[1].set_title("log(1 + precio)"); ax[1].set_xlabel("log(1+precio)")
plt.suptitle("Figura 1. Distribucion del precio original y del logaritmo del precio"); plt.tight_layout(); plt.show()

```

Figura 2. Distribución del precio según el número de amenities

Bloque 10

```

sub = df[df["amenities_count"].between(1, 4)]
gr = [sub.loc[sub["amenities_count"] == k, "price"].dropna() for k in [1, 2, 3, 4]]
fig, ax = plt.subplots(figsize=(7, 4.5))
ax.boxplot(gr, tick_labels=["1", "2", "3", "4"], showfliers=False)
ax.set_xlabel("nº de amenities"); ax.set_ylabel("precio (€)")
ax.set_title("Figura 2. Distribucion del precio segun el numero de amenities");
plt.tight_layout(); plt.show()
sub.groupby("amenities_count")["price"].agg(["count", "median"]).round(2)

```

Figura 3. Distribución de log(1 + price) por barrio agrupado (incluye «Otros»)

Bloque 11

```

orden =
df.groupby("neighbourhood_grouped")["log_price"].median().sort_values(ascending=False)
).index.tolist()
datos = [df.loc[df["neighbourhood_grouped"] == b, "log_price"].dropna() for b in
orden]
fig, ax = plt.subplots(figsize=(9.5, 5))
ax.boxplot(datos, tick_labels=orden, showfliers=False)
ax.set_ylabel("log(1 + price)"); ax.set_xlabel("barrio agrupado (incluye «Otros»)")
plt.setp(ax.get_xticklabels(), rotation=45, ha="right")
ax.set_title("Figura 3. Distribucion de log(1+price) por barrio agrupado");
plt.tight_layout(); plt.show()

```

Figura 4. Matriz de correlaciones entre variables numéricas

Bloque 12

```

vc = ["log_price", "price", "guests_included", "beds", "bathrooms",
      "review_scores_rating", "number_of_reviews", "amenities_count"]
corr = df[vc].corr(numeric_only=True)

```

```
fig, ax = plt.subplots(figsize=(8, 7)); im = ax.imshow(corr, cmap="RdBu_r", vmin=-1,
vmax=1)
ax.set_xticks(range(len(vc))); ax.set_xticklabels(vc, rotation=45, ha="right");
ax.set_yticks(range(len(vc))); ax.set_yticklabels(vc)
for i in range(len(vc)):
    for j in range(len(vc)): ax.text(j, i, f"{corr.iloc[i,j]:.2f}", ha="center",
va="center", fontsize=8)
plt.colorbar(im, fraction=0.046, pad=0.04); ax.grid(False)
plt.title("Figura 4. Matriz de correlaciones entre variables numericas");
plt.tight_layout(); plt.show()
```

6. Modelo de regresión (OLS, errores robustos HC3)

Categorías de referencia (omitidas por drop_first): barrio = ****Carabanchel****; tipo de propiedad = ****apartamento**** (Private room in apartment).

Bloque 13

```
variables = ["log_price", "guests_included", "beds", "bathrooms",
            "review_scores_rating", "number_of_reviews", "amenities_count",
            "property_type_grouped", "neighbourhood_grouped"]
df_model = pd.get_dummies(df[variables].copy(),
                        columns=["property_type_grouped",
"neighbourhood_grouped"], drop_first=True)
for c in df_model.columns:
    if df_model[c].dtype == "bool": df_model[c] = df_model[c].astype(int)
df_model = df_model.apply(pd.to_numeric, errors="coerce").replace([np.inf, -np.inf],
np.nan).dropna()

Y = df_model["log_price"]; X = sm.add_constant(df_model.drop(columns=["log_price"]))
modelo = sm.OLS(Y, X).fit(cov_type="HC3")
print(f"N = {int(modelo.nobs)}   R2 = {modelo.rsquared:.4f}   "
      f"R2 ajustado = {modelo.rsquared_adj:.4f}   F p-valor = "
      f"{modelo.f_pvalue:.2e}")
modelo.summary()
```

6.1 Tabla de coeficientes (Tabla 4)

Bloque 14

```
pd.DataFrame({"coef": modelo.params, "err_std_HC3": modelo.bse,
            "t": modelo.tvalues, "p_valor": modelo.pvalues,
            "IC_2.5%": modelo.conf_int()[0], "IC_97.5%":
modelo.conf_int()[1]}).round(4)
```

Figura 5. Coeficientes estimados con intervalos de confianza (95%)

Bloque 15

```
coefs = modelo.params.drop("const"); ic = modelo.conf_int().drop("const"); err =
(ic[1]-ic[0])/2
o = coefs.sort_values().index
lbl = [n.replace("neighbourhood_grouped_", "barrio: ")
      n.replace("property_type_grouped_", "tipo: ") for n in o]
fig, ax = plt.subplots(figsize=(8, 7))
ax.errorbar(coefs[o], range(len(o)), xerr=err[o], fmt="o", color="#3E5C76",
ecolor="gray", capsize=3)
ax.axvline(0, color="#B5485D", linestyle="--", lw=1); ax.set_yticks(range(len(o)));
ax.set_yticklabels(lbl)
```

```
ax.set_xlabel("coeficiente (efecto sobre log(1+price))"); ax.set_title("Figura 5. Coeficientes estimados con IC (95%)"); plt.tight_layout(); plt.show()
```

7. Diagnóstico del modelo

Figura 6. Valores reales vs. valores ajustados (con R^2 ajustado)

Bloque 16

```
fitted = modelo.fittedvalues
r2_fit = r2_score(Y, fitted)

# R² ajustado del ajuste (in-sample)
n = len(Y)
p = X.shape[1] - 1          # nº de predictores, sin la constante
r2_fit_adj = 1 - (1 - r2_fit) * (n - 1) / (n - p - 1)
print(f"R² del ajuste : {r2_fit:.4f}")
print(f"R² ajustado   : {r2_fit_adj:.4f}   (n={n}, p={p})")

fig, ax = plt.subplots(figsize=(6, 6))
ax.scatter(fitted, Y, s=10, alpha=0.4, color="#3E5C76")
lims = [min(fitted.min(), Y.min()), max(fitted.max(), Y.max())]; ax.plot(lims, lims,
"--", color="#B5485D", lw=1)
ax.set_xlabel("valores ajustados"); ax.set_ylabel("valores reales log(1+price)")
ax.set_title(f"Figura 6. Reales vs. ajustados (R² ajustado = {r2_fit_adj:.3f})");
plt.tight_layout(); plt.show()
```

Figura 7. Histograma de los residuos del modelo

Bloque 17

```
resid = modelo.resid
fig, ax = plt.subplots(figsize=(7, 4.5)); ax.hist(resid, bins=40, color="#5B8A72",
edgecolor="white")
ax.set_xlabel("residuo"); ax.set_ylabel("frecuencia"); ax.set_title("Figura 7. Histograma de los residuos"); plt.tight_layout(); plt.show()
```

Figura 8. QQ-plot de los residuos

Bloque 18

```
fig, ax = plt.subplots(figsize=(6, 6)); stats.probplot(resid, dist="norm", plot=ax)
ax.set_title("Figura 8. QQ-plot de los residuos"); plt.tight_layout(); plt.show()
```

7.1 Multicolinealidad (VIF) y heterocedasticidad (Breusch-Pagan)

Dos diagnósticos formales que respaldan la especificación: el VIF comprueba que las variables explicativas no están excesivamente correlacionadas entre sí, y el test de Breusch-Pagan contrasta la presencia de heterocedasticidad, que justifica el uso de errores estándar robustos HC3.

Bloque 19

```
from statsmodels.stats.outliers_influence import variance_inflation_factor
from statsmodels.stats.diagnostic import het_breuschpagan

# --- VIF (variables continuas; los VIF de dummies no son informativos) ---
```

```

continuas = ["guests_included", "beds", "bathrooms", "review_scores_rating",
             "number_of_reviews", "amenities_count"]
Xv = sm.add_constant(df_model[continuas])
vif = pd.DataFrame({
    "variable": continuas,
    "VIF": [variance_inflation_factor(Xv.values, i + 1) for i in
           range(len(continuas))]
}).round(2)
print("Factor de inflacion de la varianza (VIF):")
print(vif.to_string(index=False))
print("Regla practica: VIF < 5 indica ausencia de multicolinealidad relevante.")

# --- Breusch-Pagan (heterocedasticidad) ---
lm, lm_p, f_stat, f_p = het_breuschpagan(modelo.resid, modelo.model.exog)
print("")
print(f"Test de Breusch-Pagan: LM = {lm:.2f}, p-valor = {lm_p:.3e}")
print("p < 0,05 -> se rechaza la homocedasticidad: hay heterocedasticidad,")
print("lo que justifica el uso de errores estandar robustos (HC3).")

```

8. Evaluación predictiva fuera de muestra (con R^2 ajustado)

Partición 80/20. MAE y RMSE en escala de $\log(1+\text{precio})$.

Bloque 20

```

# X ya incluye la constante (añadida en la sección 6), no se vuelve a añadir
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2,
                                                    random_state=42)
modelo_tr = sm.OLS(Y_train, X_train).fit(cov_type="HC3")
Y_pred = modelo_tr.predict(X_test)

r2_test = r2_score(Y_test, Y_pred)

# R² ajustado sobre el conjunto de prueba
n_test = len(Y_test)
p = X.shape[1] - 1      # nº de predictores, sin la constante
r2_test_adj = 1 - (1 - r2_test) * (n_test - 1) / (n_test - p - 1)

print(f"R² (muestra completa)      : {modelo.rsquared:.4f}")
print(f"R² ajustado (modelo)        : {modelo.rsquared_adj:.4f}")
print(f"R² (entrenamiento)           : {modelo_tr.rsquared:.4f} (N train =
{len(Y_train)}))")
print(f"R² (prueba)                  : {r2_test:.4f} (N test = {n_test})")
print(f"R² ajustado (prueba)         : {r2_test_adj:.4f}")
print(f"MAE (prueba)                  : {mean_absolute_error(Y_test, Y_pred):.4f}")
print(f"RMSE (prueba)                : {np.sqrt(mean_squared_error(Y_test,
Y_pred)):.4f}")

```

Figura 9. Valores reales vs. valores predichos (conjunto de prueba)

Bloque 21

```

val = f"{r2_test_adj:.3f}"
fig, ax = plt.subplots(figsize=(6.5, 6))
ax.scatter(Y_test, Y_pred, s=14, alpha=0.5, color="#3E5C76", label="Observaciones")
lims = [min(Y_test.min(), Y_pred.min()), max(Y_test.max(), Y_pred.max())]
ax.plot(lims, lims, "--", color="#B5485D", lw=1.3, label="Predicción perfecta
(45°)")

# Línea de tendencia (ajuste lineal de los puntos)

```

```

z = np.polyfit(Y_test, Y_pred, 1)
poly = np.poly1d(z)
xs = np.linspace(Y_test.min(), Y_test.max(), 100)
ax.plot(xs, poly(xs), color="#5B8A72", lw=2, label="Línea de tendencia")

ax.set_xlabel("valores reales log(1+price)")
ax.set_ylabel("valores predichos")
ax.set_title(f"Figura 9. Capacidad predictiva del modelo (R2 ajustado = {r2_test_adj:.3f})")
ax.legend(loc="upper left", fontsize=8)
plt.tight_layout(); plt.show()

```

9. Análisis de robustez: tipo de propiedad desagregado (11 categorías)

El modelo principal agrupa el tipo de propiedad en tres categorías (apartamento / casa / otro). Como comprobación de robustez se reestima usando las **once** categorías originales de `property_type` mediante variables dummy. Se contrasta que el efecto de `bathrooms` se mantiene estable y que la mejora de ajuste al desagregar es marginal, lo que respalda la agrupación.

Bloque 22

```

variables_alt = ["log_price", "guests_included", "beds", "bathrooms",
                 "review_scores_rating", "number_of_reviews", "amenities_count",
                 "property_type", "neighbourhood_grouped"]
m_alt = pd.get_dummies(df[variables_alt].copy(),
                      columns=["property_type", "neighbourhood_grouped"],
                      drop_first=True)
for c in m_alt.columns:
    if m_alt[c].dtype == "bool":
        m_alt[c] = m_alt[c].astype(int)
m_alt = m_alt.apply(pd.to_numeric, errors="coerce").replace([np.inf, -np.inf],
np.nan).dropna()

Y_alt = m_alt["log_price"]
X_alt = sm.add_constant(m_alt.drop(columns=["log_price"]))
modelo_alt = sm.OLS(Y_alt, X_alt).fit(cov_type="HC3")

print("Comparacion de especificaciones del tipo de propiedad")
print(f" Principal (3 categorias: apto/casa/otro) : R2 = {modelo.rsquared:.4f}
R2 ajustado = {modelo.rsquared_adj:.4f}")
print(f" Alternativa (11 categorias property_type) : R2 =
{modelo_alt.rsquared:.4f} R2 ajustado = {modelo_alt.rsquared_adj:.4f}")
print("")
print(f" Coef. bathrooms principal = {modelo.params['bathrooms']:+.4f}")
print(f" Coef. bathrooms alternativa = {modelo_alt.params['bathrooms']:+.4f}")
print("El signo y la magnitud de bathrooms se mantienen: el resultado es robusto,")
print("y desagregar en 11 categorias apenas mejora el ajuste -> se prefiere la
agrupacion.")

```

Nota técnica: el notebook completo se entrega junto con este documento como archivo complementario (TFG_Analytics_FINAL.ipynb). La reproducibilidad de los resultados está garantizada mediante el uso de `random_state=42` en la partición train/test del conjunto de datos.