



Universidad
Andrés Bello®

CLASIFICACIÓN DE VIVIENDAS MEDIANTE REGRESIÓN LOGÍSTICA

Solemne 1: Minería de datos

Integrantes: Jaime Arriagada - Sergio Villegas - Amanda Arias - Aylin Herrera

Fecha: 28/09/2025

Profesor: John Ríos Griego

IMPORTANCIA DEL ANÁLISIS

Este estudio busca predecir la categoría de precio de una vivienda (Cara/Barata) en función de sus características estructurales y ubicación, aplicando técnicas de Ciencia de Datos.

El proceso se compone de tres etapas cruciales:

- Preprocesamiento
- Análisis exploratorio (EDA)
- Modelación mediante Regresión Logística



METAS DEL ANÁLISIS

El propósito central de este análisis es desarrollar un modelo de clasificación preciso y confiable que pueda predecir si el precio de una vivienda se sitúa por encima o por debajo de la mediana del mercado, creando así una dicotomía clara entre propiedades "Caras" y "Baratas". Para alcanzar esta meta general, se han establecido objetivos específicos que guían la investigación.



DESCRIPCIÓN DEL CONJUNTO DE DATOS



El conjunto de datos "Housing" constituye la base empírica de este estudio, conteniendo información detallada de 545 observaciones, cada una representando una vivienda diferente, descrita a través de 13 variables iniciales.



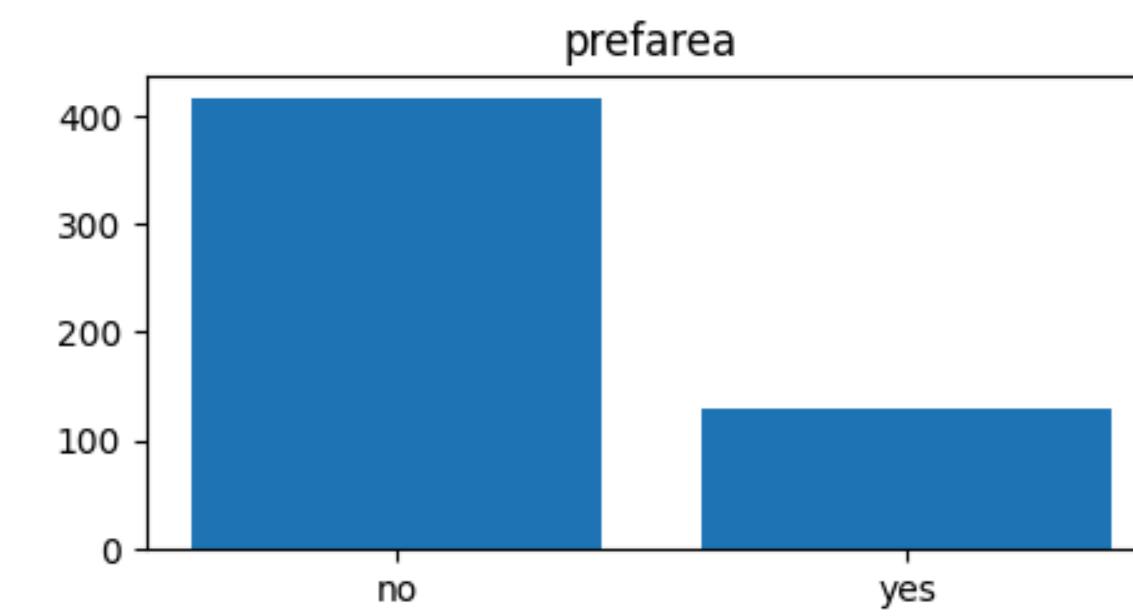
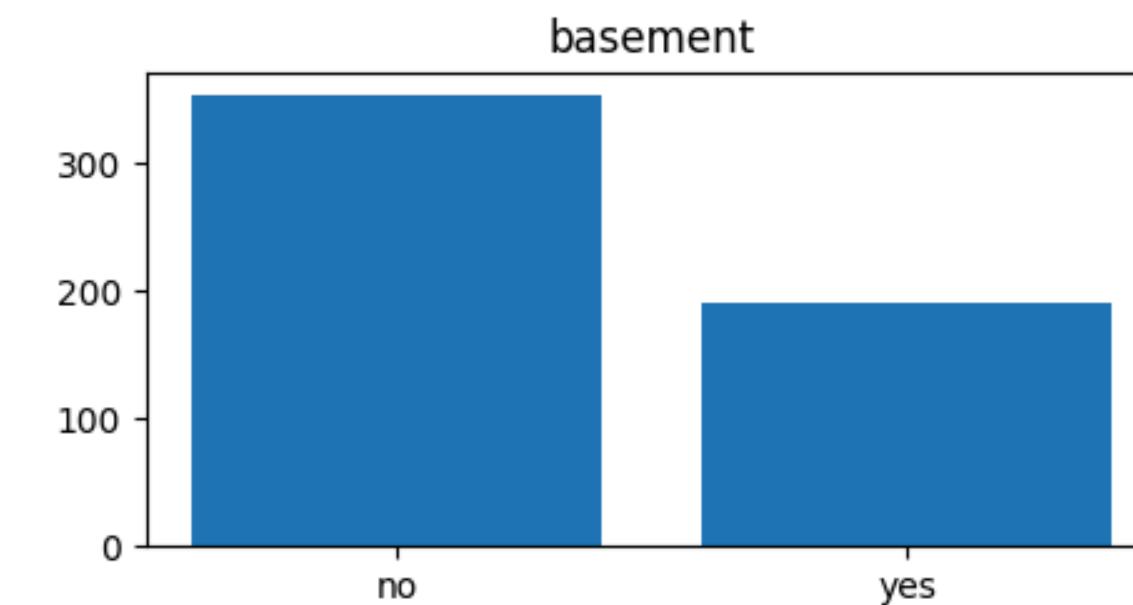
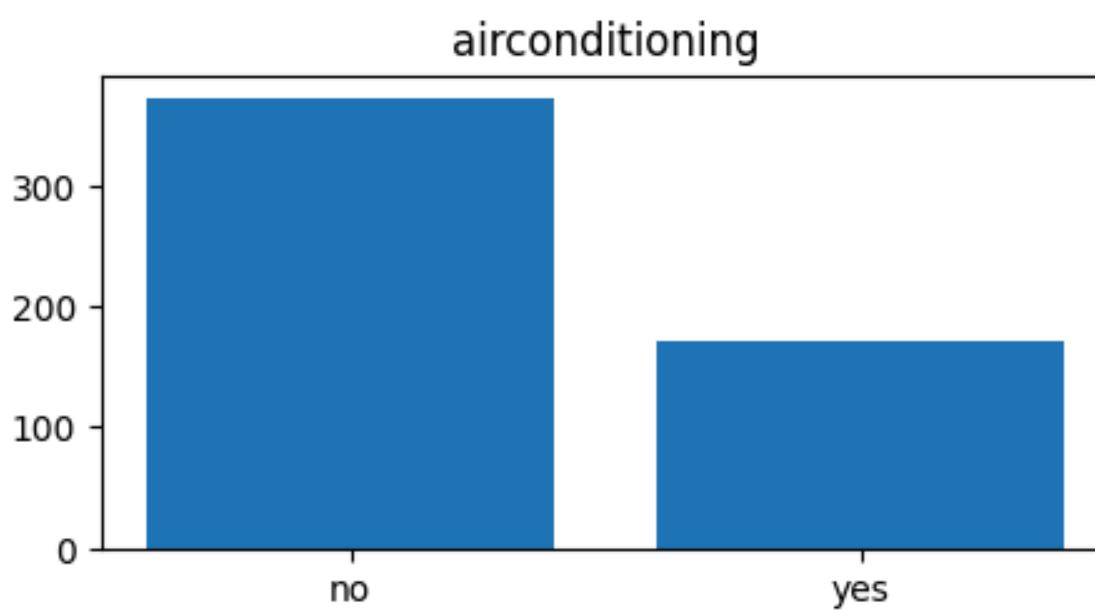
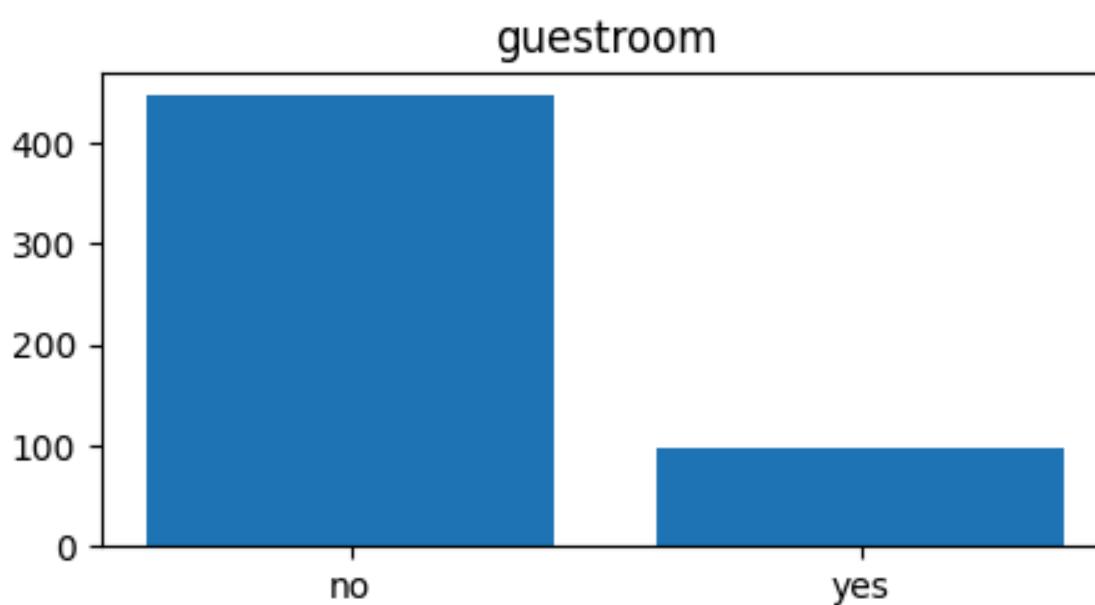
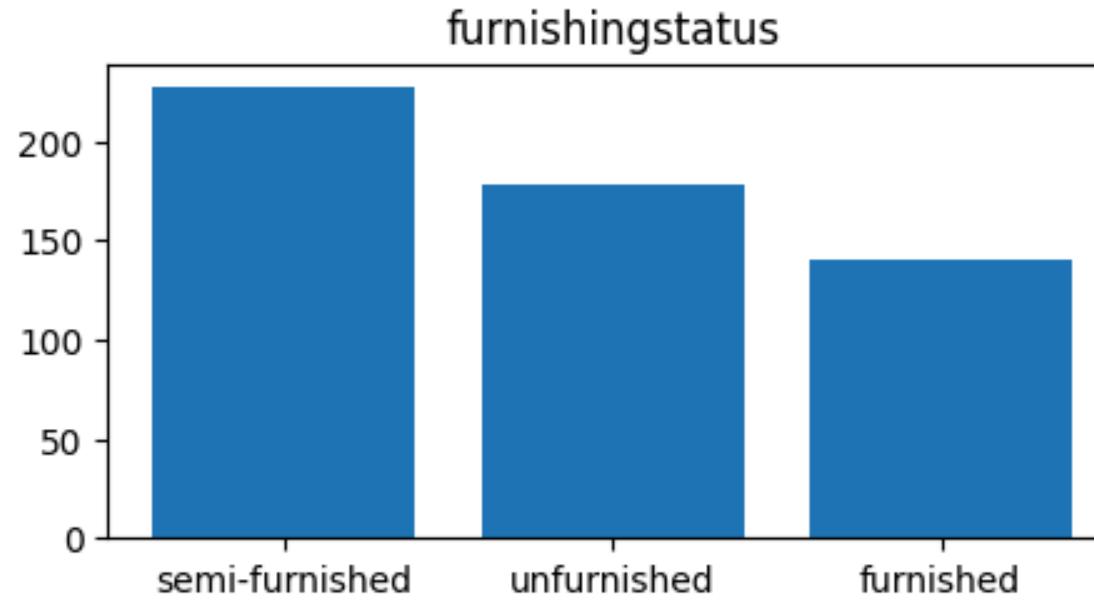
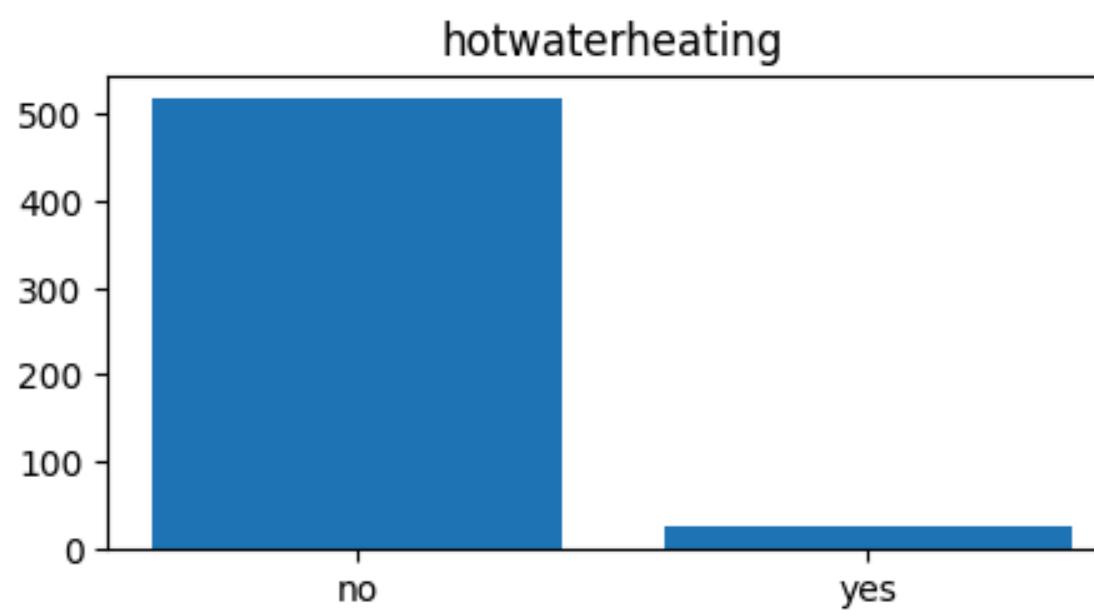
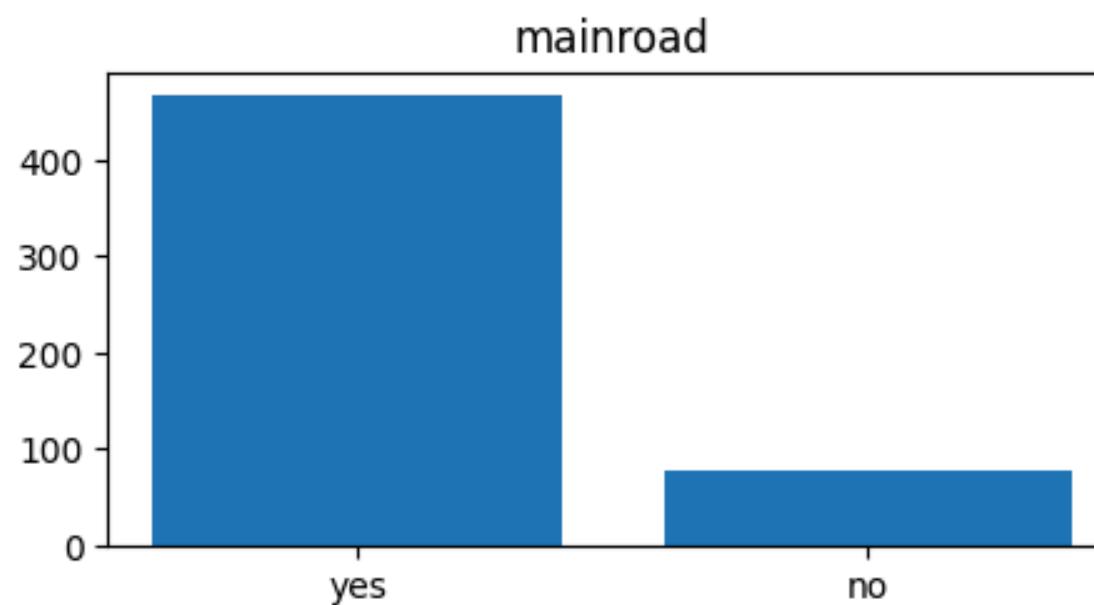
La variable objetivo original (price) fue transformada en una variable categórica binaria (PriceCategory) utilizando la mediana como punto de corte, una estrategia que asegura un balance perfecto entre las dos clases a predecir.



Las variables se clasifican en numéricas y categóricas. Entre las numéricas encontramos price (precio), area (área construida en pies cuadrados), bedrooms (número de dormitorios), bathrooms (número de baños), stories (número de pisos) y parking (número de plazas de estacionamiento).

ANÁLISIS GRÁFICO DE VARIABLES CATEGÓRICAS

El análisis de la distribución de la variable objetivo PriceCategory revela un balanceo óptimo entre las clases, con una distribución casi perfectamente equitativa entre las categorías "Cara" y "Barata". Esta distribución balanceada es ideal para el proceso de modelado, ya que elimina el riesgo de crear un algoritmo sesgado hacia una clase mayoritaria, lo que garantiza que las métricas de evaluación como la precisión y el recall sean representativas del desempeño real del modelo en ambas clases por igual.



RESUMEN DE VARIABLES NUMÉRICAS

El resumen estadístico de las variables numéricas ofrece una instantánea cuantitativa de sus comportamientos centrales y de dispersión. Los datos muestran una amplia variabilidad en variables clave como area y price, indicando la presencia de propiedades con características muy diversas dentro del mercado estudiado. Las medidas de tendencia central (media, mediana) y dispersión (desviación estándar, valores mínimos y máximos) para cada variable numérica proporcionan una base sólida para entender la escala y variabilidad de los datos con los que el modelo deberá trabajar.

RESUMEN DE VARIABLES NUMÉRICAS

3. *Medidas estadísticas de variables numéricas:*

	price	area	bedrooms	bathrooms	stories	parking
count	5.450000e+02	545.000000	545.000000	545.000000	545.000000	545.000000
mean	4.766729e+06	5150.541284	2.965138	1.286239	1.805505	0.693578
std	1.870440e+06	2170.141023	0.738064	0.502470	0.867492	0.861586
min	1.750000e+06	1650.000000	1.000000	1.000000	1.000000	0.000000
25%	3.430000e+06	3600.000000	2.000000	1.000000	1.000000	0.000000
50%	4.340000e+06	4600.000000	3.000000	1.000000	2.000000	0.000000
75%	5.740000e+06	6360.000000	3.000000	2.000000	2.000000	1.000000
max	1.330000e+07	16200.000000	6.000000	4.000000	4.000000	3.000000

Precio de Viviendas:

- **Precio promedio:** \$4.77 millones
- **Rango amplio:** desde \$1.75M hasta \$13.3M
- **50% de las propiedades** están entre \$3.43M y \$5.74M

Área Construida:

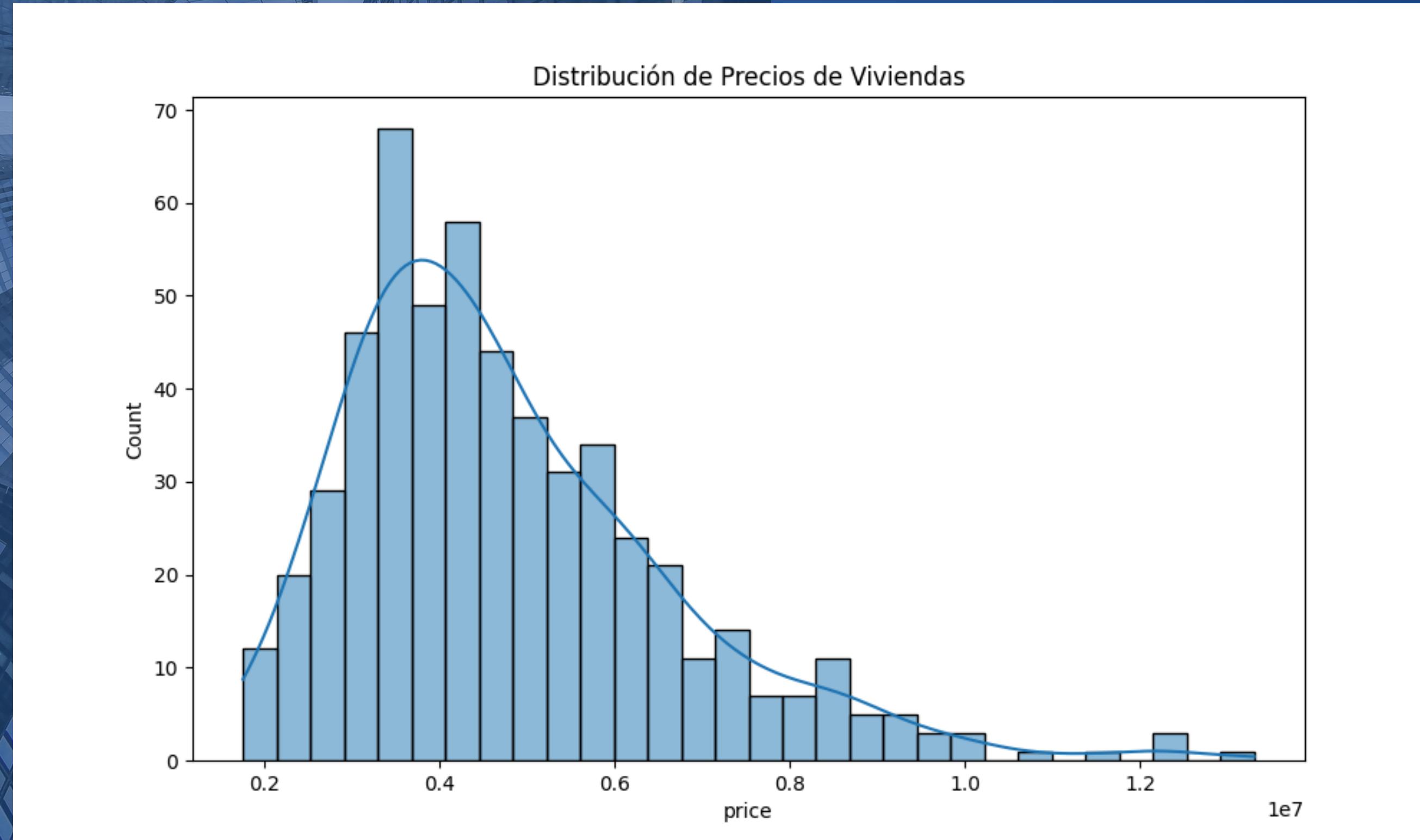
- **Área promedio:** 5,150 unidades (asumido pies²)
- **Extremos significativos:** desde 1,650 hasta 16,200 unidades
- **Mediana:** 4,600 unidades, indicando distribución asimétrica

Características Estructurales:

- **Dormitorios:** promedio de 3, mayoría entre 2-3
- **Baños:** promedio 1.3, con 50% teniendo solo 1 baño
- **Pisos:** mayoría de 1-2 pisos (75% del mercado)
- **Estacionamientos:** 70% tiene 0-1 plaza

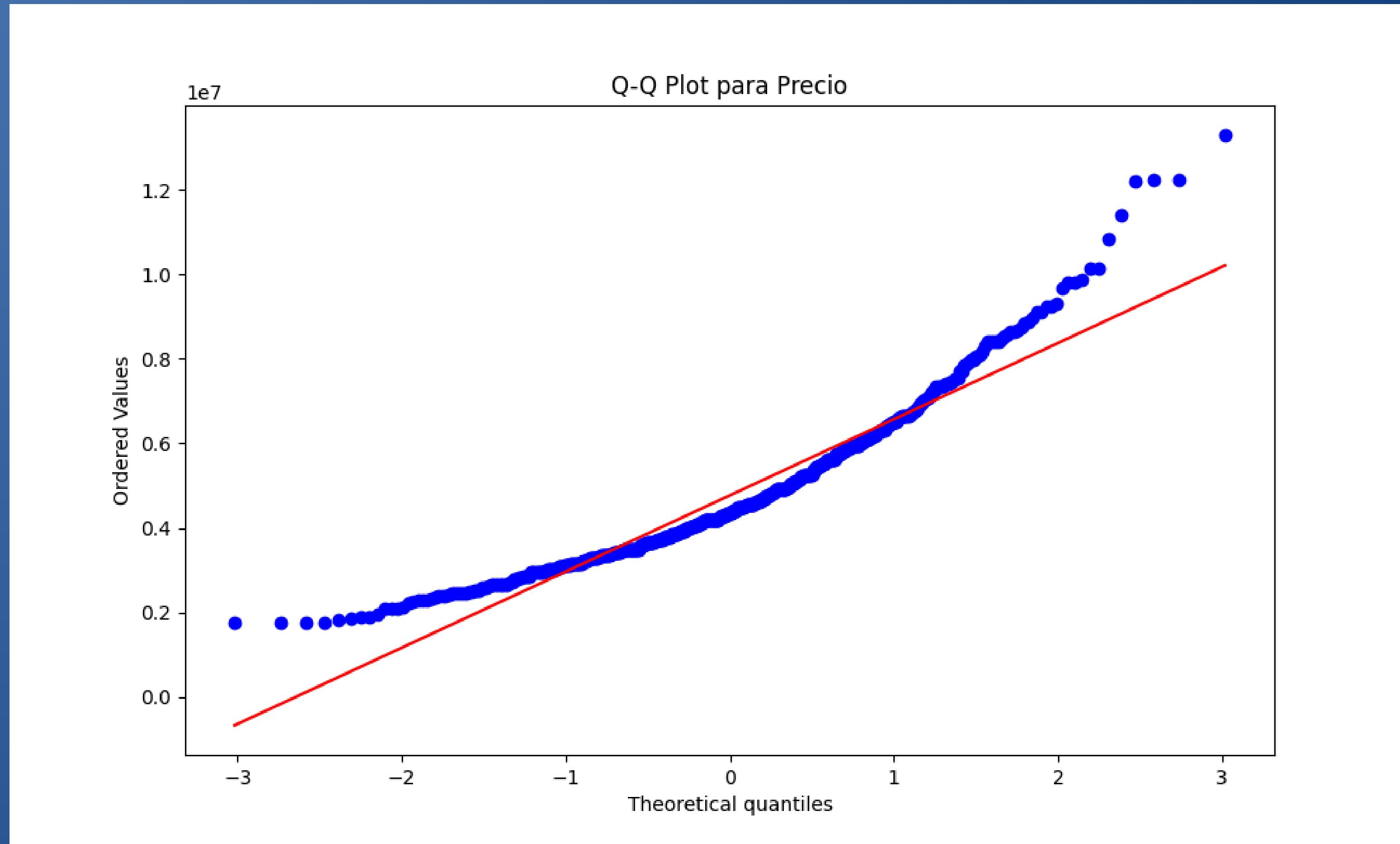
HISTOGRAMA Y DENSIDAD

Existe una distribución claramente asimétrica hacia la derecha, lo que se conoce como sesgo positivo.



ANÁLISIS DE NORMALIDAD (SHAPIRO-WILK)

La evaluación formal de normalidad para la variable area mediante la prueba de Shapiro-Wilk arrojó un p-valor extremadamente significativo ($p < 2.2e-16$), lo que lleva a rechazar de manera contundente la hipótesis nula de normalidad. Este hallazgo estadístico valida la asimetría observada visualmente en el histograma y justifica plenamente la elección de un modelo de Regresión Logística, que no depende del supuesto de normalidad en la distribución de las variables predictoras, a diferencia de otras técnicas como la regresión lineal.



Shapiro-Wilk Test: estadístico=0.9216311651574154, p-value=3.154903020052395e-16

IDENTIFICACIÓN DE DATOS ATÍPICOS (OUTLIERS)

BOXPLOT

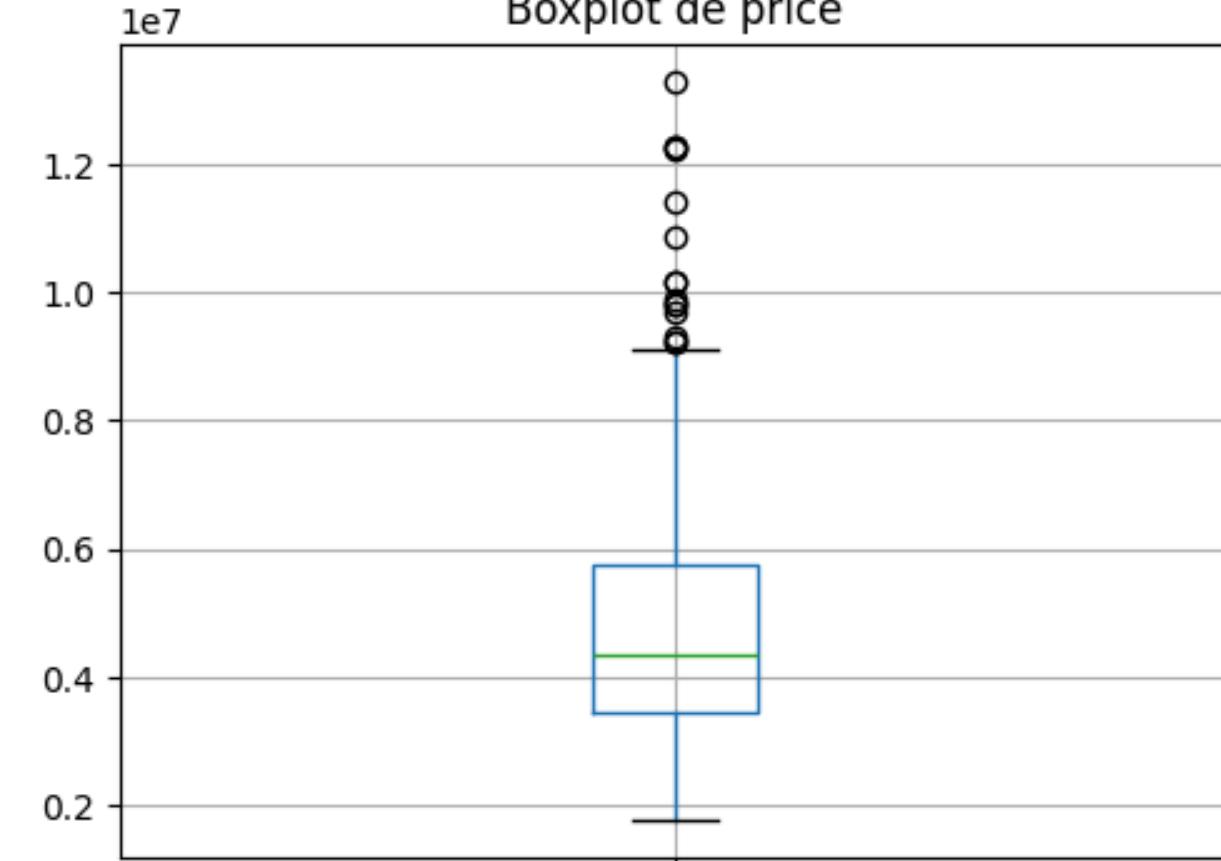
La inspección del precio mediante un diagrama de caja identifica la presencia de múltiples valores atípicos en el extremo superior de la distribución.

OUTLIERS

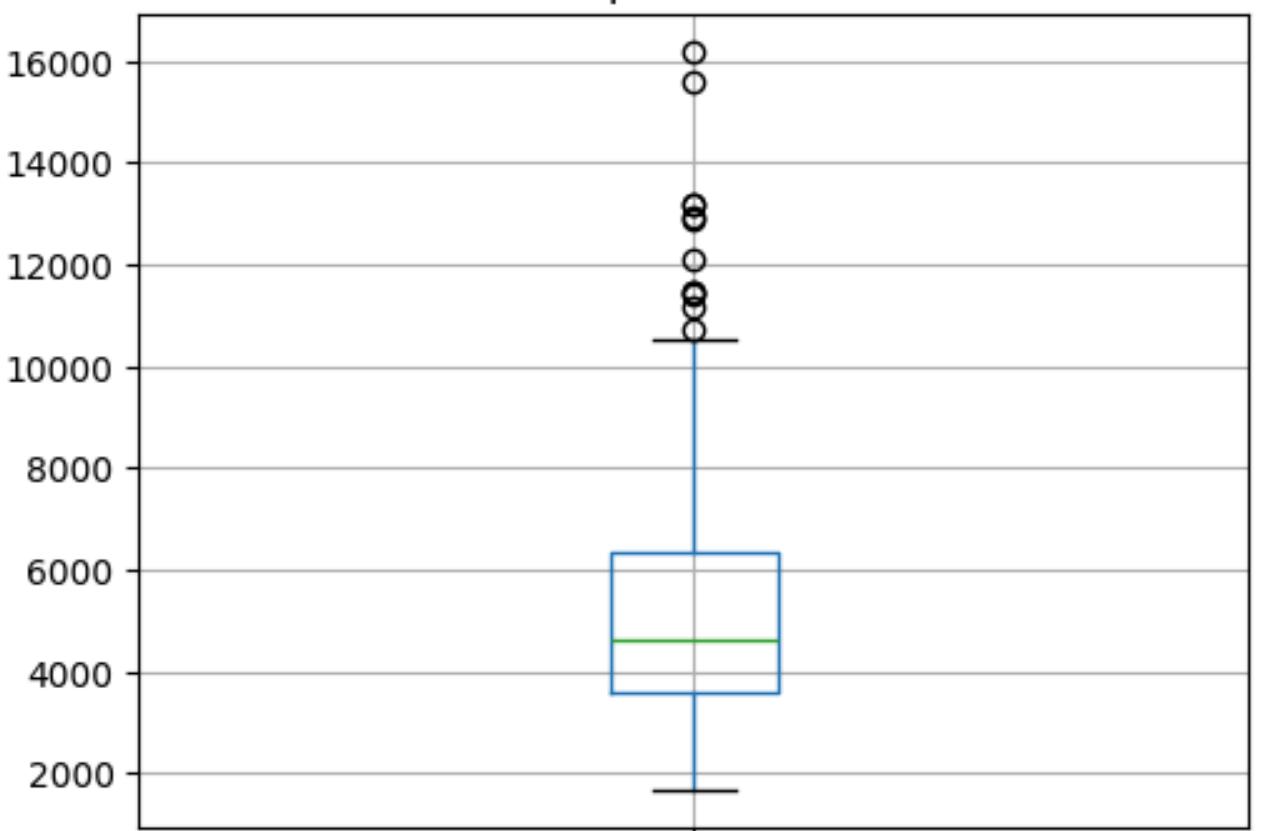
Estos outliers corresponden a propiedades que, dentro de este mercado específico, se consideran de precio excepcionalmente alto.

Se tomó la decisión consciente de no eliminar estos valores del análisis, ya que representan observaciones válidas y genuinas que forman parte de la realidad del mercado inmobiliario que se busca modelar, y su exclusión podría crear un modelo menos representativo.

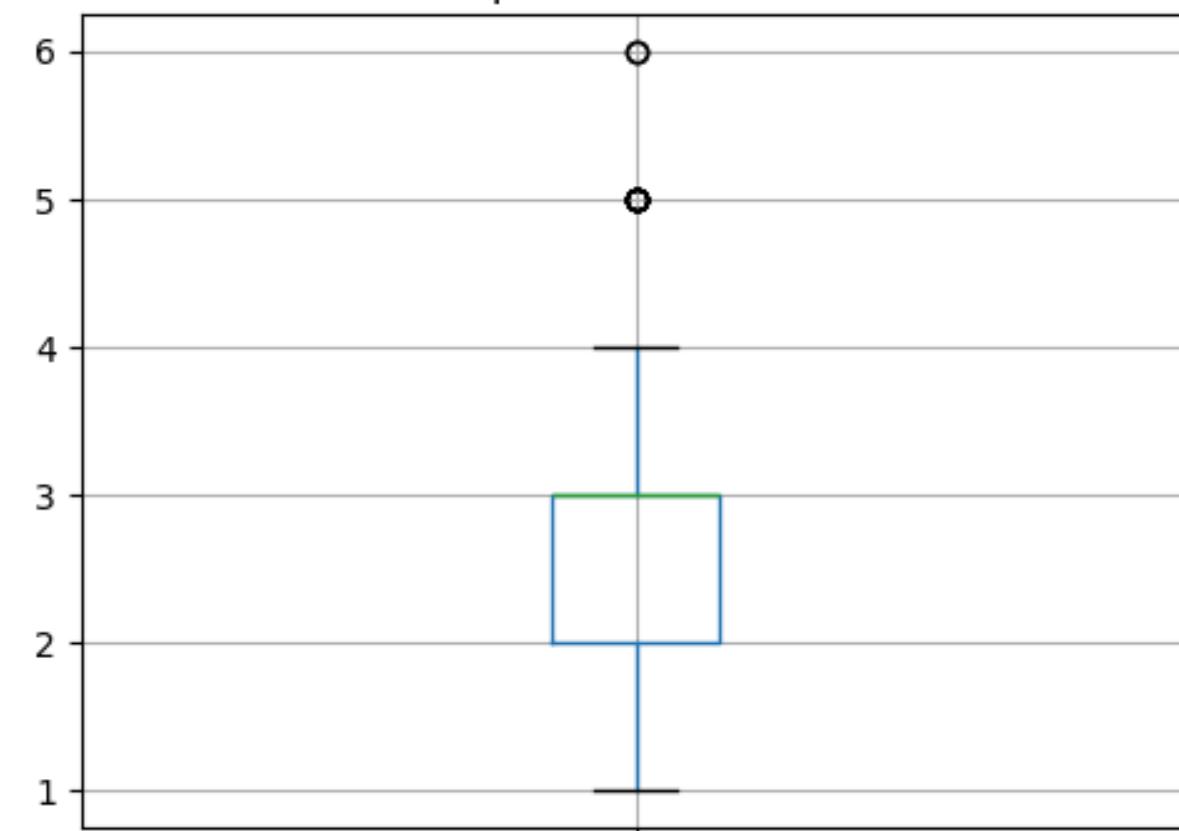
Boxplot de price



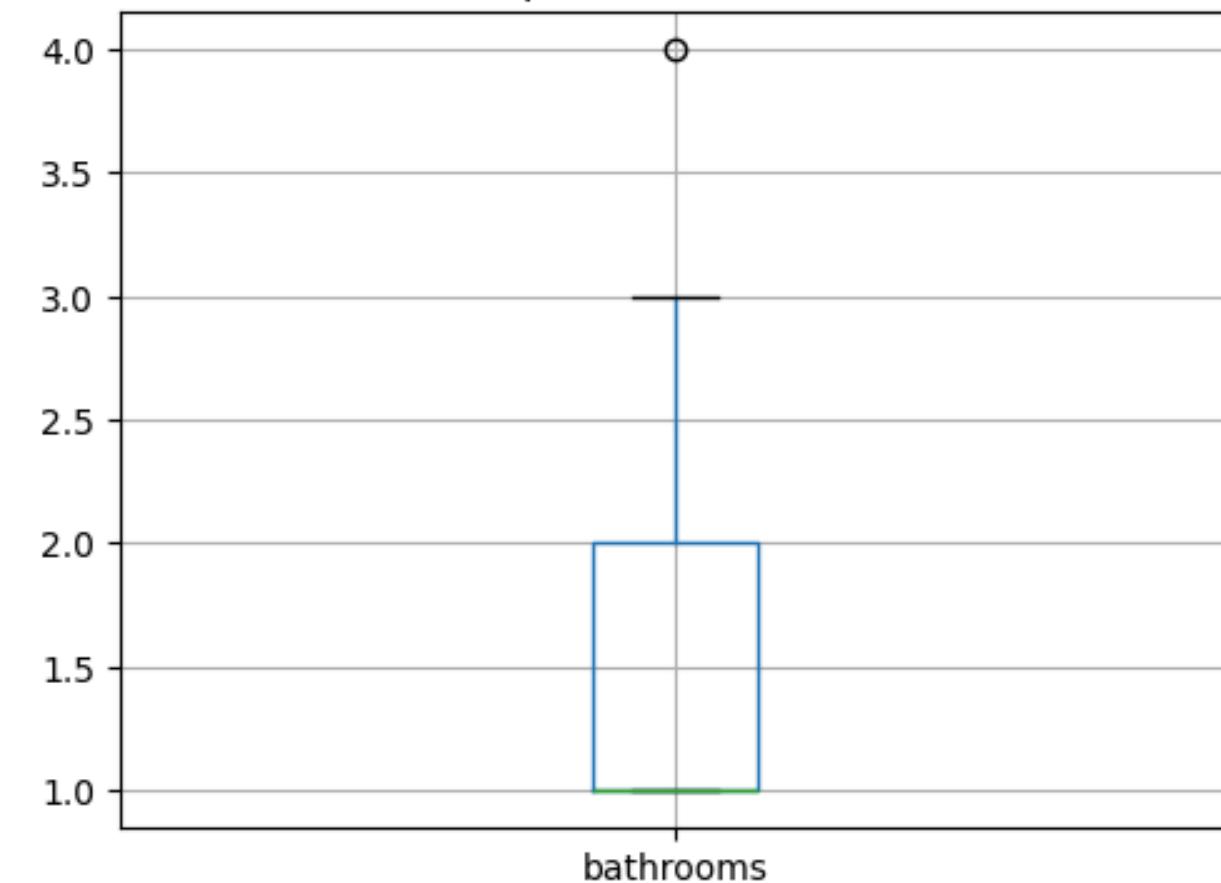
Boxplot de area



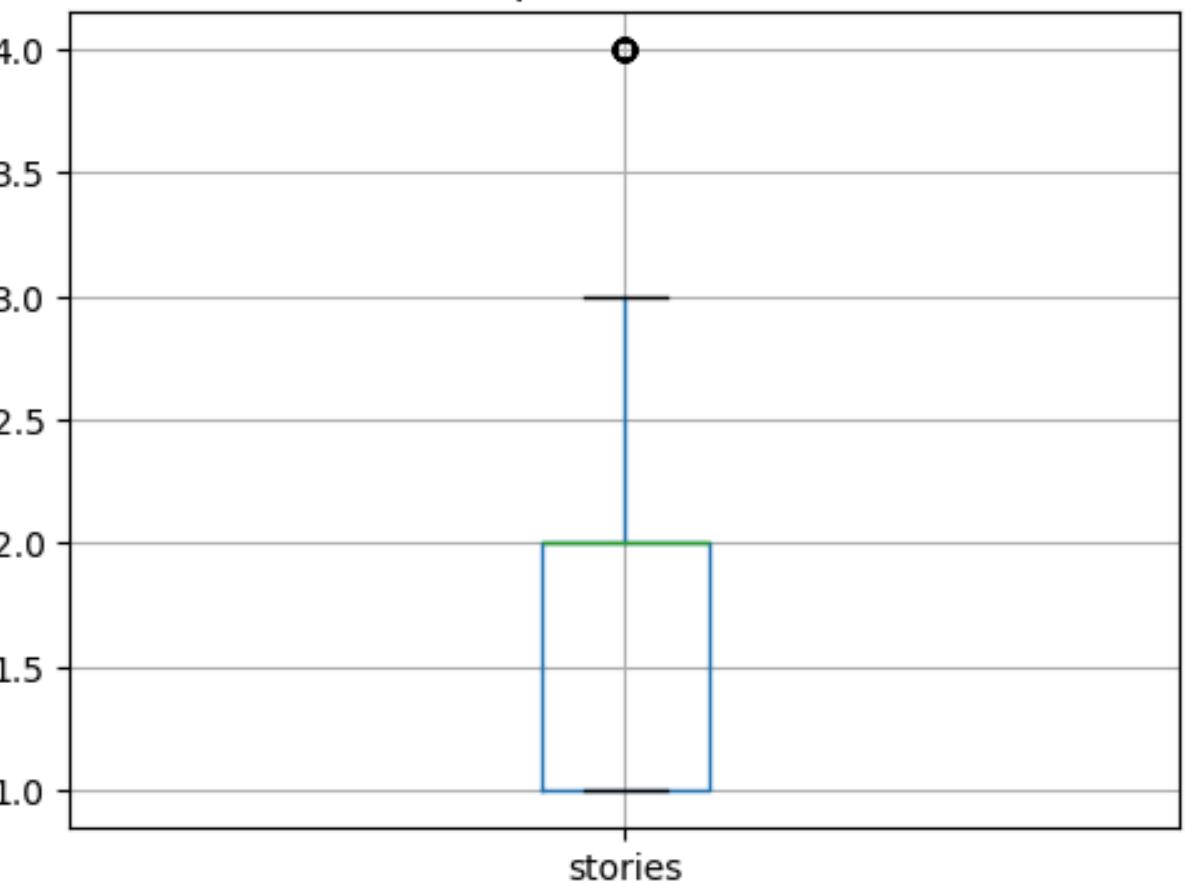
Boxplot de bedrooms



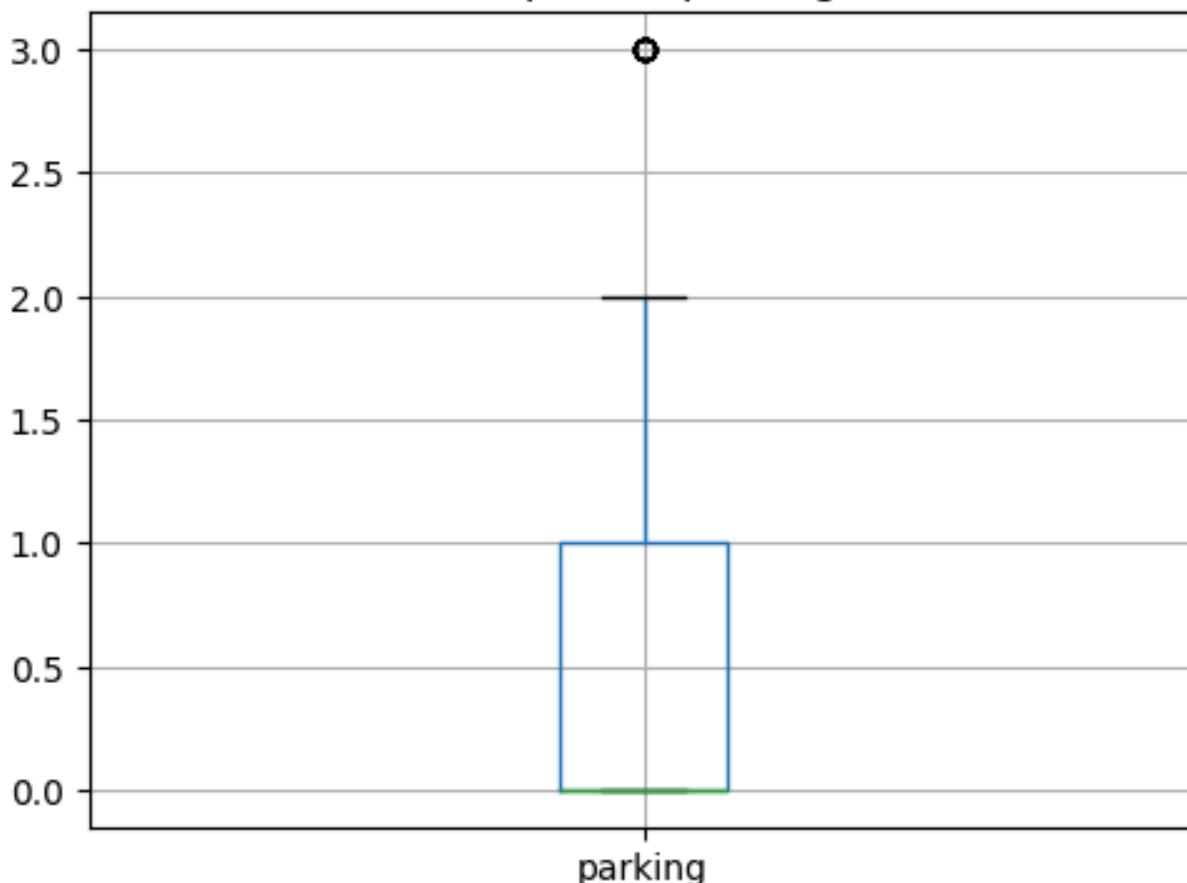
Boxplot de bathrooms



Boxplot de stories



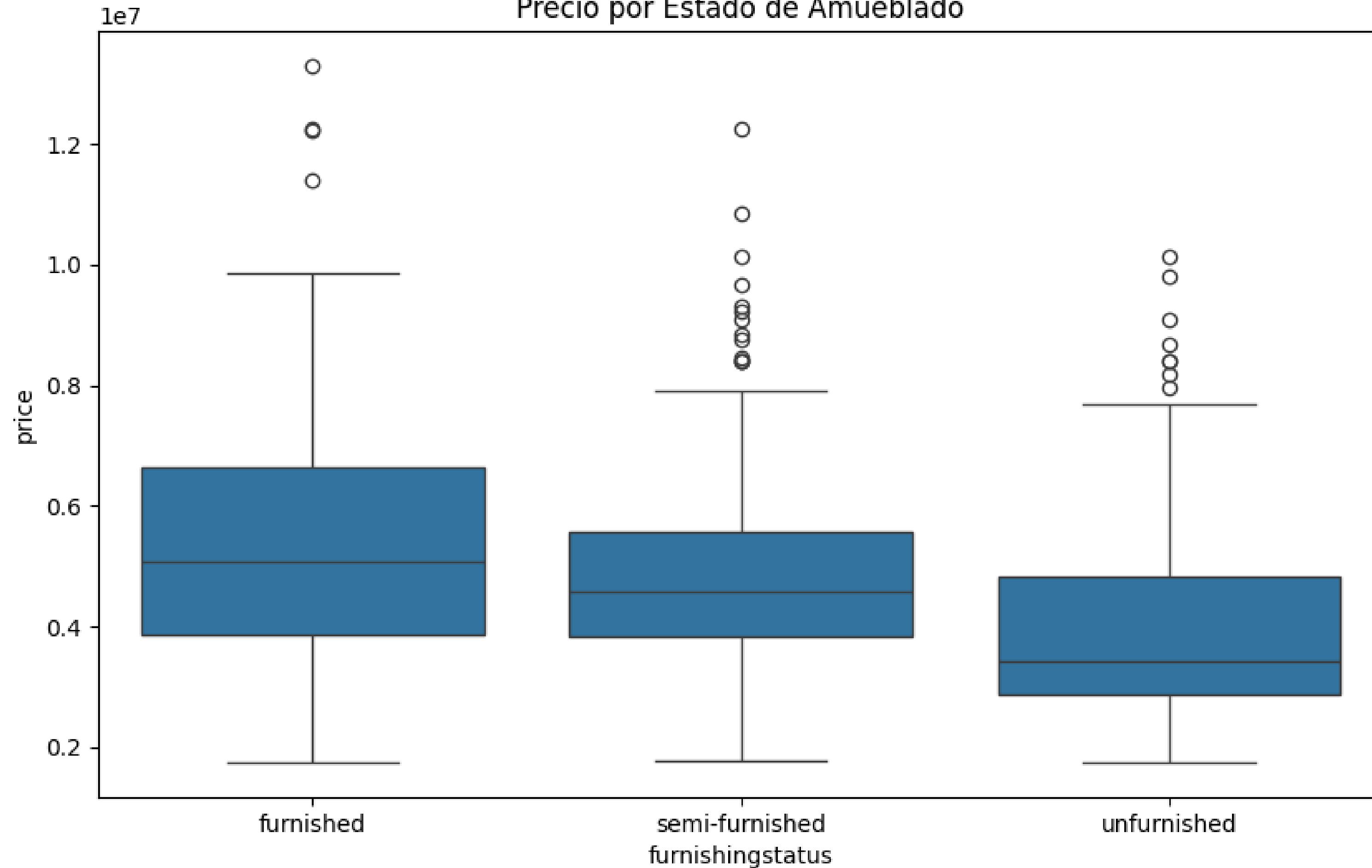
Boxplot de parking



RELACIÓN VARIABLE NUMÉRICA VS. CATEGÓRICA

El análisis de la relación entre el área (variable numérica) y la categoría de precio (variable categórica) mediante un boxplot comparativo revela una diferencia sustancial y esperable entre los dos grupos. Las viviendas categorizadas como "Caras" presentan una mediana de área significativamente mayor, y su distribución general se desplaza hacia valores más altos en comparación con las viviendas "Baratas". Esta clara separación visual sugiere de manera contundente que el área de la vivienda será uno de los predictores más importantes en el modelo de clasificación.

Precio por Estado de Amueblado



Análisis de Datos Faltantes

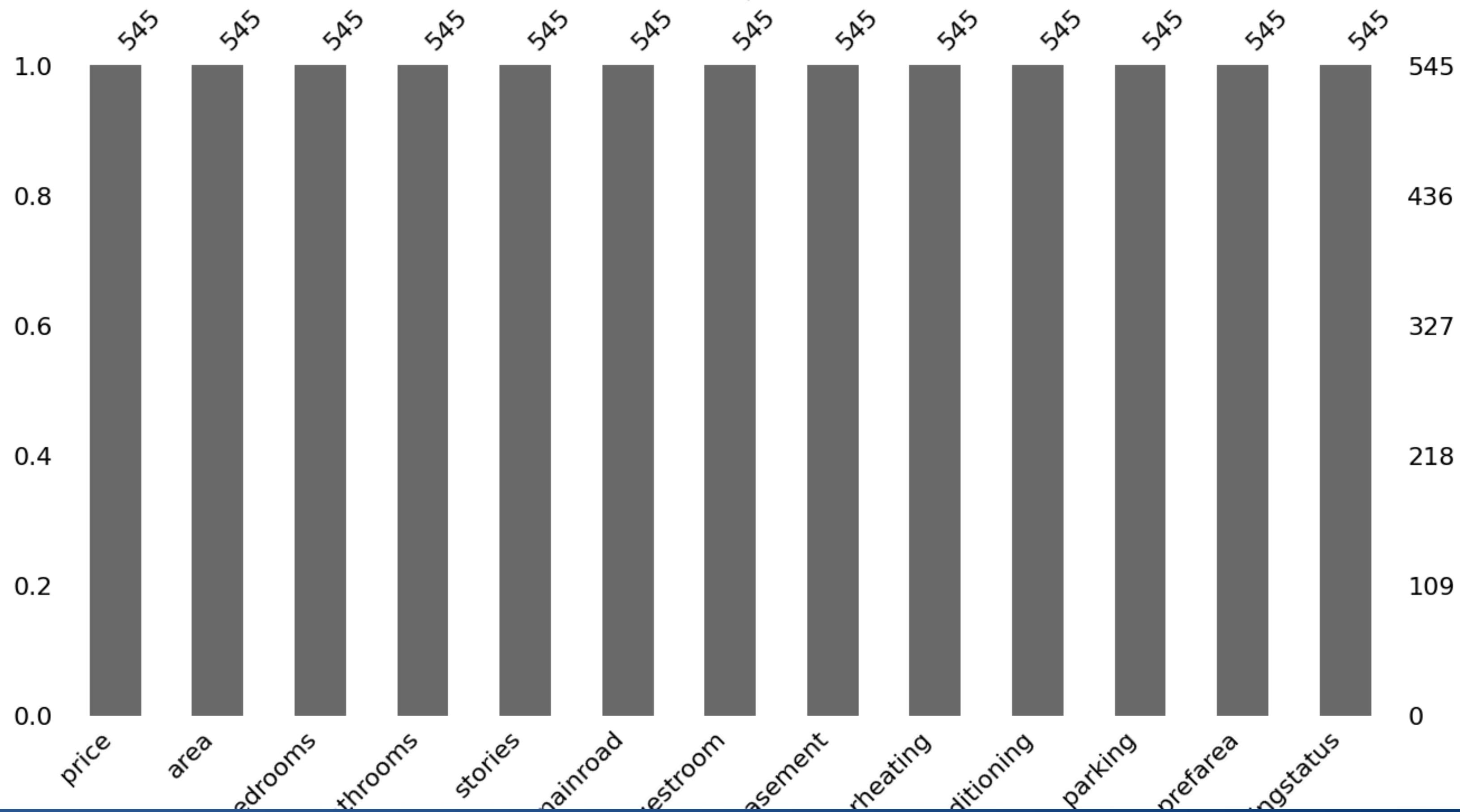
Un aspecto afortunado en la preparación de este dataset fue la completa ausencia de valores faltantes (NA) en todas las variables y observaciones. Esta integridad de datos simplificó enormemente la etapa crítica de preprocessamiento, eliminando la necesidad de implementar técnicas de imputación o eliminación de datos, lo que a su vez preserva el tamaño original de la muestra y evita introducir posibles sesgos asociados a los métodos de manejo de datos missing.

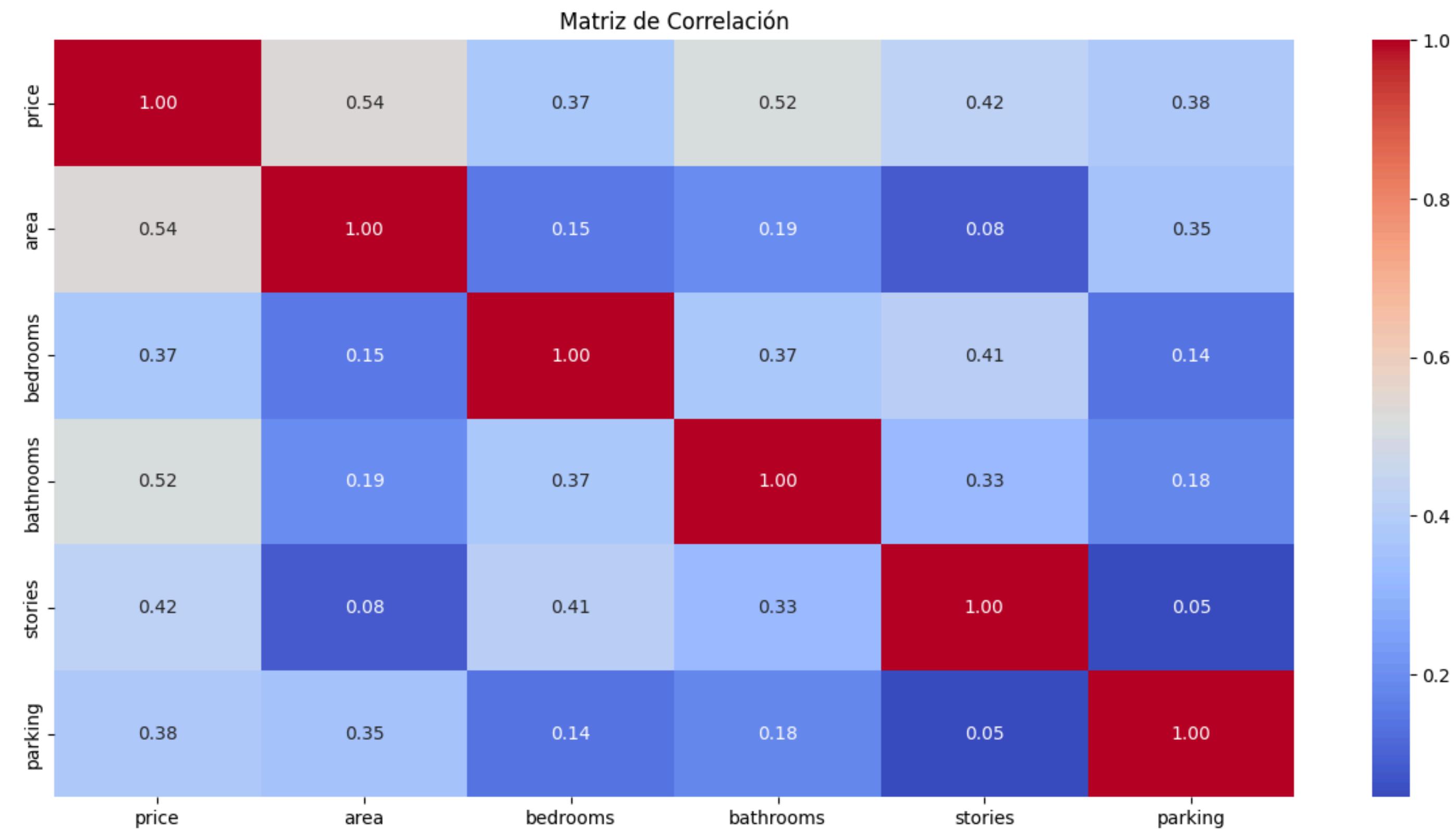
Matriz de Correlación

La matriz de correlación calculada para las variables numéricas revela la existencia de relaciones lineales positivas entre varios atributos. La correlación más fuerte y esperable se observa entre price y area, confirmando que el tamaño es un driver fundamental del precio. Se encontraron además correlaciones moderadas entre bathrooms y bedrooms, y entre stories y price. Es importante destacar que la ausencia de correlaciones extremadamente altas (superiores a 0.9) indica que no existen problemas severos de multicolinealidad que pudieran distorsionar los coeficientes del modelo de regresión logística.



Datos Faltantes por Columna





RESUMEN DE MODELADO Y EVALUACIÓN

Modelo Inicial (Todas las variables):

Variables significativas identificadas: area, bathrooms, stories, airconditioningyes, prefareayes.

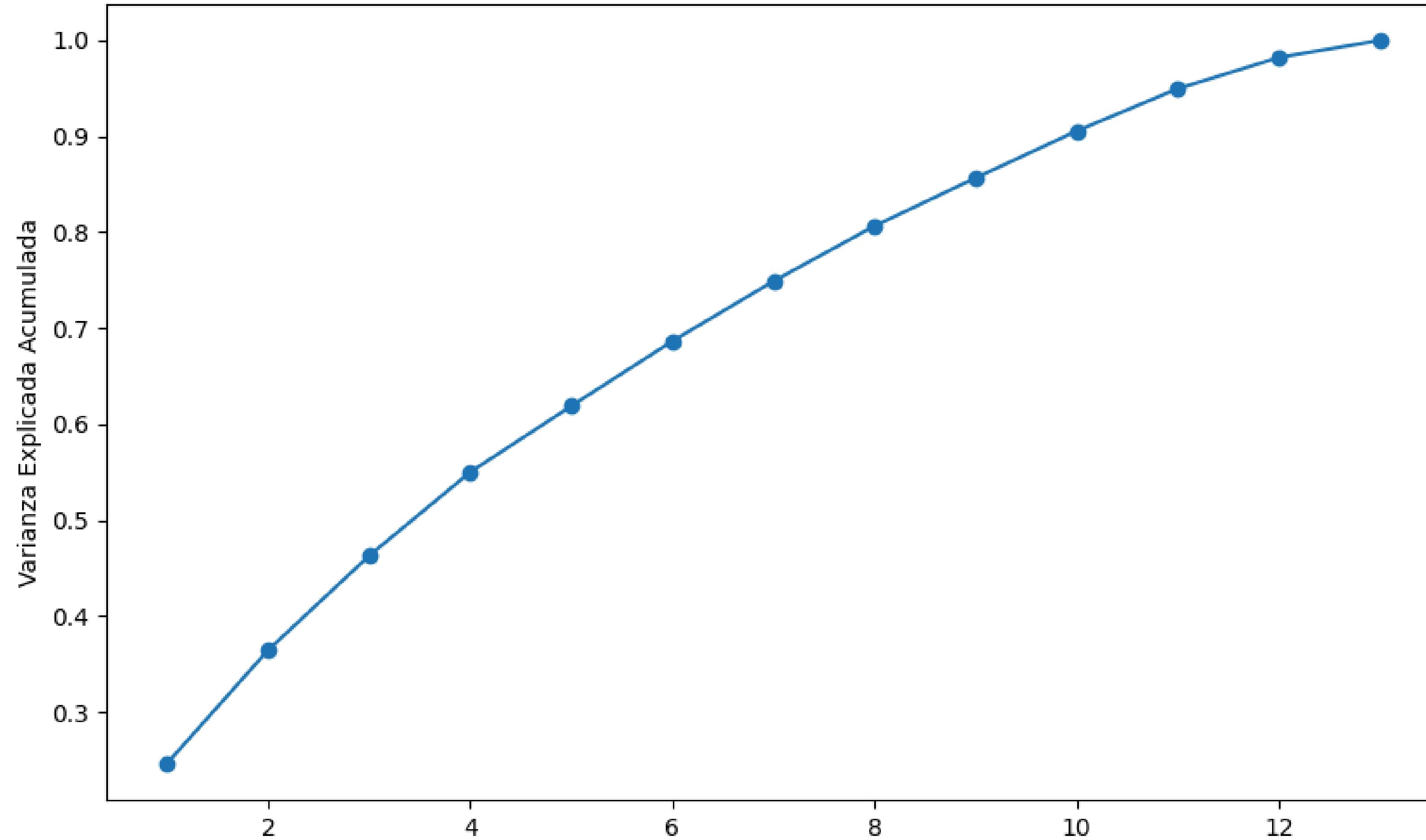
Selección de Variables (Método Backward):

Se aplicó backward stepwise con criterio AIC, obteniendo un modelo más simple y eficiente sin perder capacidad predictiva.

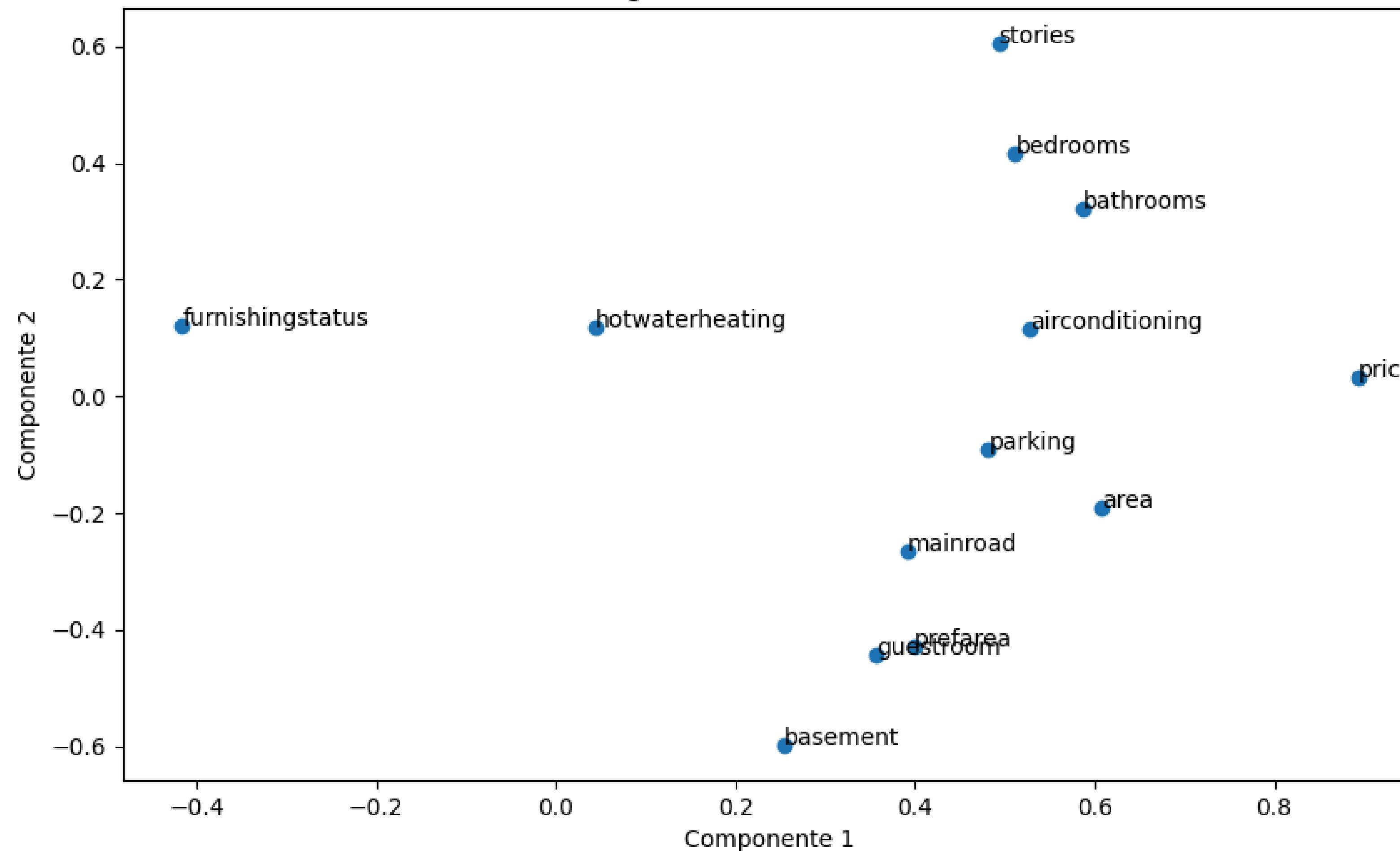
Evaluación del Modelo Final:

El modelo final, evaluado en un conjunto de prueba independiente, demostró un alto rendimiento en todas las métricas clave. La combinación de una elevada exactitud, precisión, sensibilidad y un valor AUC destacado confirma su capacidad robusta y confiable para clasificar viviendas según su categoría de precio.

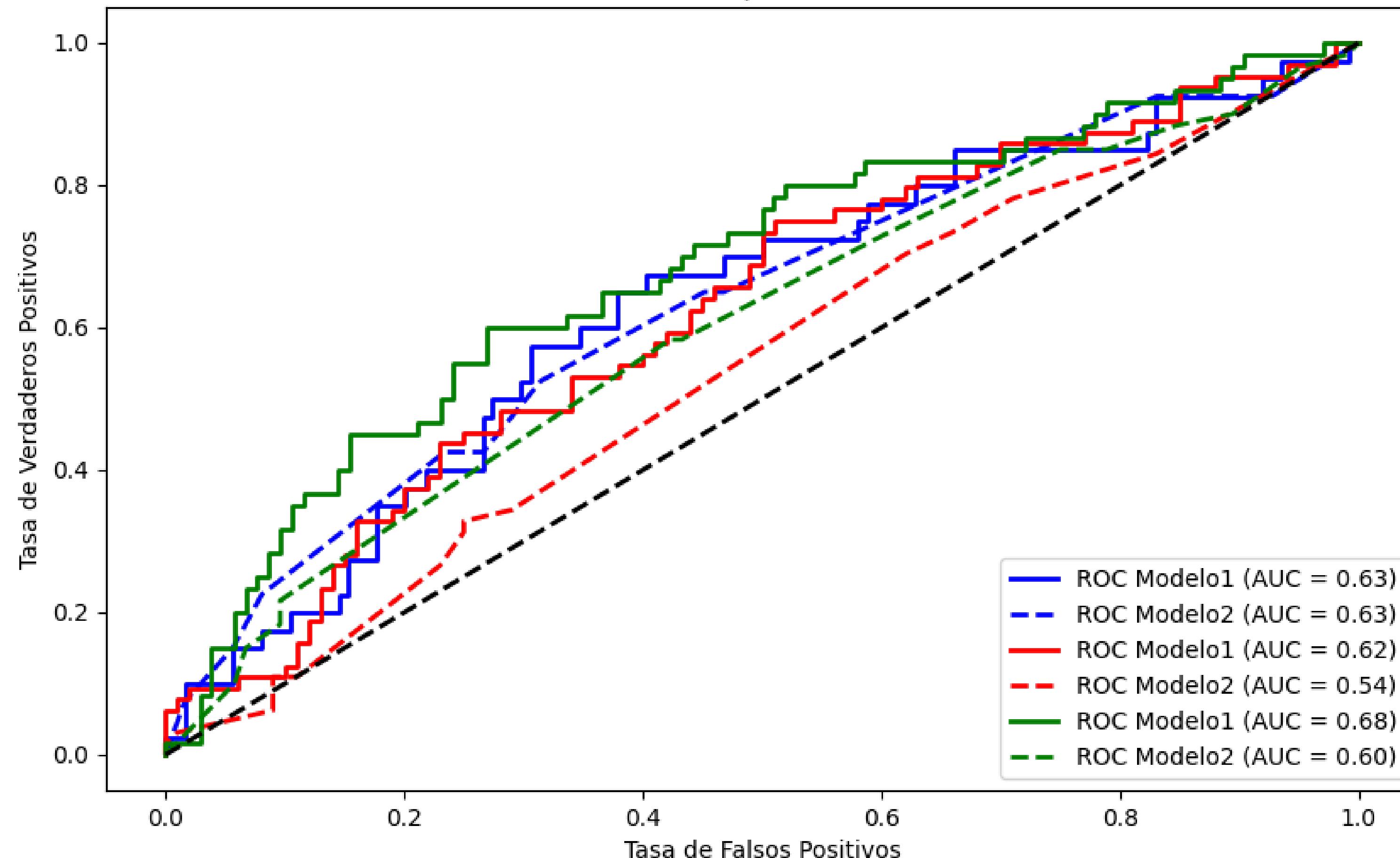
Varianza Explicada por Componentes



Cargas de Variables en ACP



Curvas ROC para Multiclas



11. Modelo 1 - Todas las variables:

Accuracy: 0.4878048780487805

Matriz de confusión:

```
[[ 4 30  6]
 [ 3 50 11]
 [ 4 30 26]]
```

Reporte de clasificación:

	precision	recall	f1-score	support
0	0.36	0.10	0.16	40
1	0.45	0.78	0.57	64
2	0.60	0.43	0.50	60
accuracy			0.49	164
macro avg	0.47	0.44	0.41	164
weighted avg	0.49	0.49	0.45	164

12. Modelo 2 - Backward Selection:

Accuracy: 0.4329268292682927

Matriz de confusión:

```
[[ 6 31  3]
 [ 4 56  4]
 [ 3 48  9]]
```

Reporte de clasificación:

	precision	recall	f1-score	support
0	0.46	0.15	0.23	40
1	0.41	0.88	0.56	64
2	0.56	0.15	0.24	60
accuracy			0.43	164
macro avg	0.48	0.39	0.34	164
weighted avg	0.48	0.43	0.36	164

13. Comparación de Modelos:

	Modelo	Accuracy	Precision	Recall	F1-Score
0	Todas las variables	0.487805	0.487289	0.487805	0.447240
1	Backward Selection	0.432927	0.480242	0.432927	0.361508

>>>> Conclusion

Este estudio permitió concluir que las características físicas inherentes a una vivienda, como su área construida, el número de baños y la cantidad de pisos, junto con comodidades específicas como la presencia de aire acondicionado y su ubicación en una zona preferencial, constituyen predictores sólidos y estadísticamente significativos para determinar su categorización de precio. El modelo de Regresión Logística demostró ser una herramienta altamente efectiva y adecuada para esta tarea de clasificación binaria, entregando resultados confiables y accionables.



>>>>



¿Qué aprendimos de los datos?

Aprendimos que el área construida, el número de baños, la cantidad de pisos, la presencia de aire acondicionado y la ubicación en zona preferencial son predictores significativos para clasificar el precio de una vivienda. La regresión logística resultó ser un modelo efectivo y interpretable para esta tarea.

¿Qué limitaciones encontramos?

El modelo no considera variables externas como factores económicos o tendencias temporales. Además, la presencia de outliers aunque reales, puede afectar la generalización del modelo en mercados con precios más homogéneos.

¿Qué mejoraríamos en un próximo análisis?

Incluiríamos más variables contextuales (como cercanía a servicios o índice de criminalidad), probaríamos otros algoritmos (como Random Forest o XGBoost) y aplicaríamos técnicas de validación cruzada para mejorar la robustez del modelo.



BIBLIOGRAFIA

Chen, H. (2024). House price prediction based on machine learning model. En M. R. Mohyuddin & N. A. D. IDE (Eds.), Proceeding of the 2024 International Conference on Diversified Education and Social Development (DESD 2024) (pp. 133-142). Advances in Social Science, Education and Humanities Research, 899. Atlantis Press. https://doi.org/10.2991/978-2-38476-346-7_18

Çilgin, C., Gökßen, Y., & Gökçen, H. (2023). The effect of outlier detection methods in real estate valuation with machine learning. İzmir Sosyal Bilimler Dergisi, 5(1), 9–20. <https://doi.org/10.47899/ijss.1270433>

Fu, Y. (2024). A comparative study of house price prediction using linear regression and random forest models. *Highlights in Science, Engineering and Technology*, 107, 96–103.

Gnat, S. (2021). Impact of categorical variables encoding on property mass valuation. *Procedia Computer Science*, 192, 3542–3550. <https://doi.org/10.1016/j.procs.2021.09.127>

Luzuriaga Jaramillo, H. A., Espinosa Pinos, C. A., Haro Sarango, A. F., & Ortiz Román, H. D. (2023). Histograma y distribución normal: Shapiro-Wilk y Kolmogorov Smirnov aplicado en SPSS. *LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades*, 4(4), 596–607. <https://doi.org/10.56712/latam.v4i4.1242>

Rowland López, F. (2023). Un modelo econométrico para determinar el valor de venta de proyectos inmobiliarios en la ciudad de Santiago. *Horizontes Empresariales*, 19(1), 28–50.



GRACIAS