Análisis de precios de la vivienda en Madrid

La Memoria Técnica

TEAM – M 02- EDA

Natalia Sánchez-Horneros Jiménez, Marcos Martínez Santiago, Neha Malhotra

Índice

1. Introducción

Para nuestro proyecto, realizamos un análisis exhaustivo de las 11.826 viviendas y 14 variables del mercado inmobiliario de Madrid para identificar los factores clave que influyen en los precios de las propiedades. Después de estudiar variables o factores como el tamaño y tipo de la propiedad, el distrito, la variación de precios, la zona habitable, el número de habitaciones y baños y la disponibilidad de servicios, tienen una influencia significativa en el valor de las propiedades.

El objetivo principal es identificar patrones relevantes en los precios y sus variables explicativas.

- Otros objetivos para investigar a través de estos estudios son:
- Analizar la distribución de los precios de vivienda.
- Evaluar la relación entre precio y características del inmueble.
- Detectar tendencias y posibles segmentaciones del mercado.

2. Hipótesis / Preguntas a responder

Hipótesis principal: Los barrios centrales tienen un precio por metro cuadrado significativamente más alto que los barrios periféricos.

Otras hipótesis:

- Hipótesis 2: Las propiedades reformadas tienen un precio por metro cuadrado al menos un 15% superior al de las propiedades no reformadas.
- Hipótesis 3: Los anuncios de viviendas en plantas altas muestran un precio más alto en comparación con las plantas bajas.
- Hipótesis 4: Las viviendas situadas a menos de 300 metros de una estación de metro tienen un precio por metro cuadrado más alto.

3. Exploración de datos

Nuestro conjunto de datos ha sido extraído de Idealista de Kaggle explorar distribuciones de precios (univariantes con histogramas/diagramas de caja), relaciones con características como tamaño/habitaciones (bivariados con dispersión/regplots) y, potencialmente, usar gráficos avanzados (diagramas de par, mapas de calor) para encontrar ideas para modelar, demostrando habilidades en ciencia de datos en el análisis de precios de propiedades.

El análisis exploratorio de datos (EDA) permite identificar patrones, relaciones y anomalías, facilitando una mejor comprensión del comportamiento del mercado.

Analizamos las primeras características del dataset para conocer:

- Número total de registros y columnas
- Tipos de datos por variable
- Estadísticos descriptivos básicos

Este análisis nos ayuda a identificar posibles errores, valores atípicos o columnas que necesiten ser transformadas.

```
def explo_datos():
    # Forma del dataset: (filas, columnas), tipo de los datos y descripción
    df.shape
    df.info()
    df.describe(include="all")
```

4. Distribución de la variable individual

En este parte se analizará cada variable del dataset de forma individual para comprender su distribución, valores característicos y posibles anomalías.

Análisis de variables numéricas

Vamos a ver la distribución de las variables numéricas más relacionadas con el precio:

```
def variables_numericas():
    df[["PrecioActual", "precio_m2", "metros", "habitaciones", "baños"]].describe()

    # Analizamos la distribución de las variables numéricas para identificar tendencias y valores atípicos
    plt.figure(figsize=(12,6))
    df[["PrecioActual", "precio_m2", "metros"]].hist(bins=30, figsize=(12,6))
    plt.suptitle("Distribución variables numéricas")

    # Guardado del gráfico en carpeta img
    plt.savefig("src/img/hist_numeric.png", dpi=300, bbox_inches="tight")

    plt.show()
    plt.close()
```

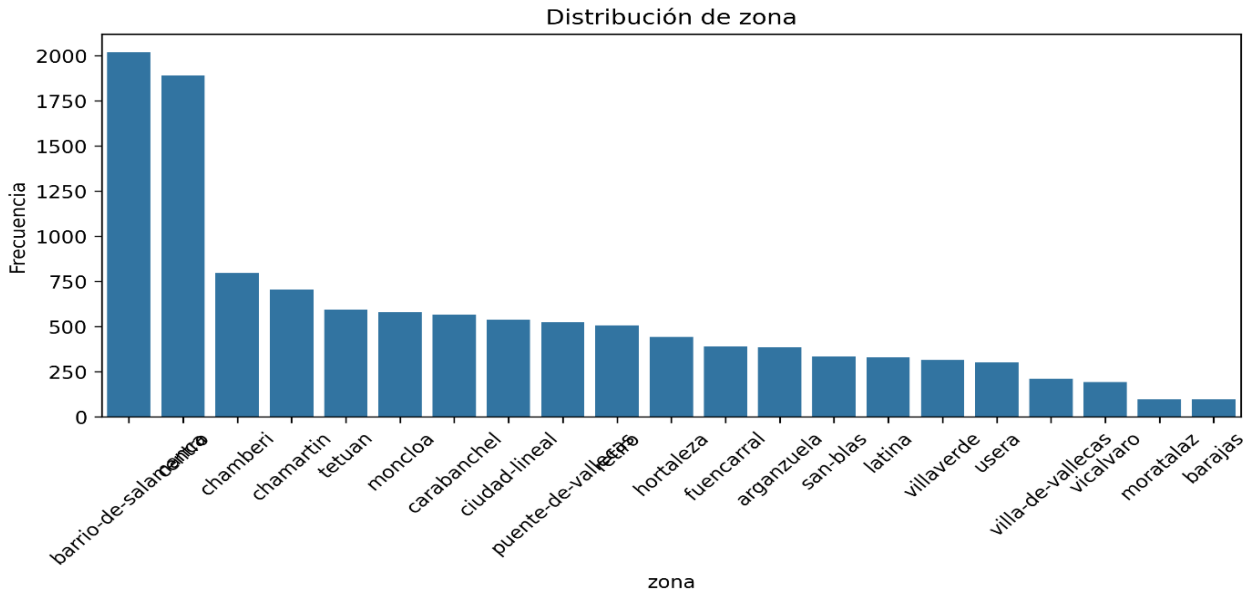
- Precio total del inmueble
- Precio por metro cuadrado
- Metros cuadrados

Variable	**Lo que te dice**	**Patrón**
PrecioActual	Precio actual de las propiedades	La mayoría son de precio medio-bajo, pocos muy caros
precio_m2	Precio por m^2	La mayoría en rangos medios, algunos agudos
metros	Tamaño de la propiedad	La mayoría pequeña, pocas grandes

5. Análisis univariante (Análisis de variables categóricas)

5.1. Zona

```python
def analisis_variables_catogoricas():
    # Guardamos las Variables categóricas a analizar
    cat_vars = ["zona", "localizacion", "ascensor"]

    for var in cat_vars:
        plt.figure(figsize=(10,4))
        sns.countplot(data=df, x=var, order=df[var].value_counts().index)
        plt.xticks(rotation=45)
        plt.title(f"Distribución de {var}")
        plt.xlabel(var)
        plt.ylabel("Frecuencia")

        # Guardado del gráfico correspondiente
        plt.savefig(f"src/img/cat_{var}.png", dpi=300, bbox_inches="tight")

        plt.show()
        plt.close()
```



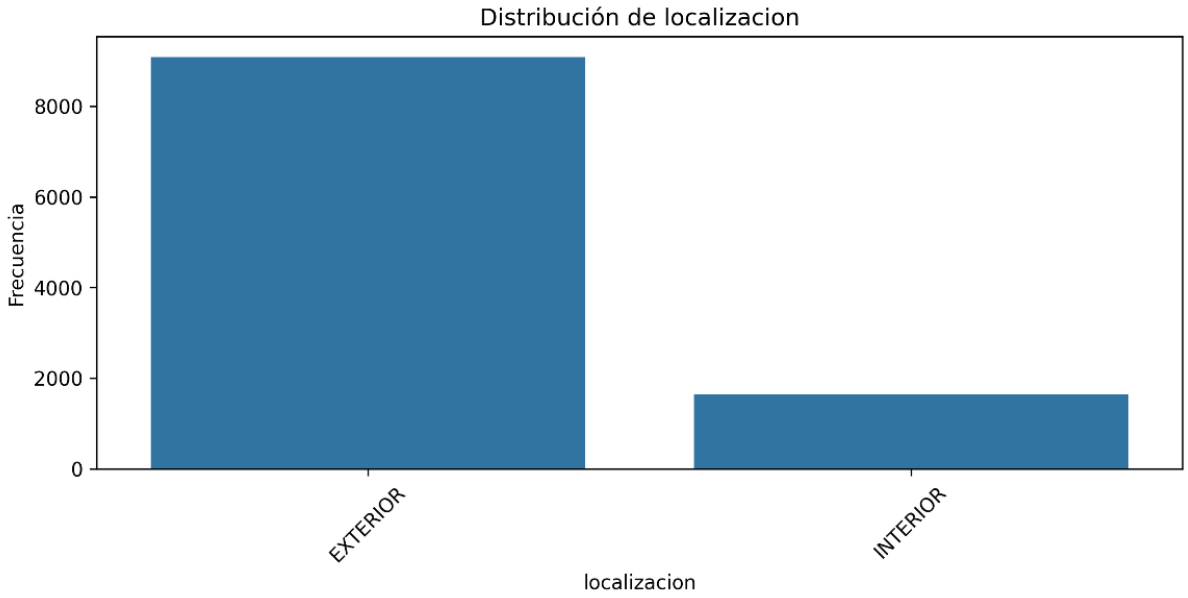
Es un gráfico de barras que muestra la distribución de una variable categórica "zona" (barrios o zonas):

- Eje X (zona): Diferentes categorías (vecindarios) en tu conjunto de datos.
- Eje Y (Frecuencia): Cuántas propiedades pertenecen a cada zona.

Cada barra representa una zona, y la altura de la barra te indica cuántos puntos de datos (por ejemplo, propiedades) hay en esa zona.

- Las zonas con alta frecuencia pueden darte estadísticas más fiables.
- Las zonas con baja frecuencia pueden requerir un manejo cuidadoso (por ejemplo, agrupación, muestreo).

5.2. Localización



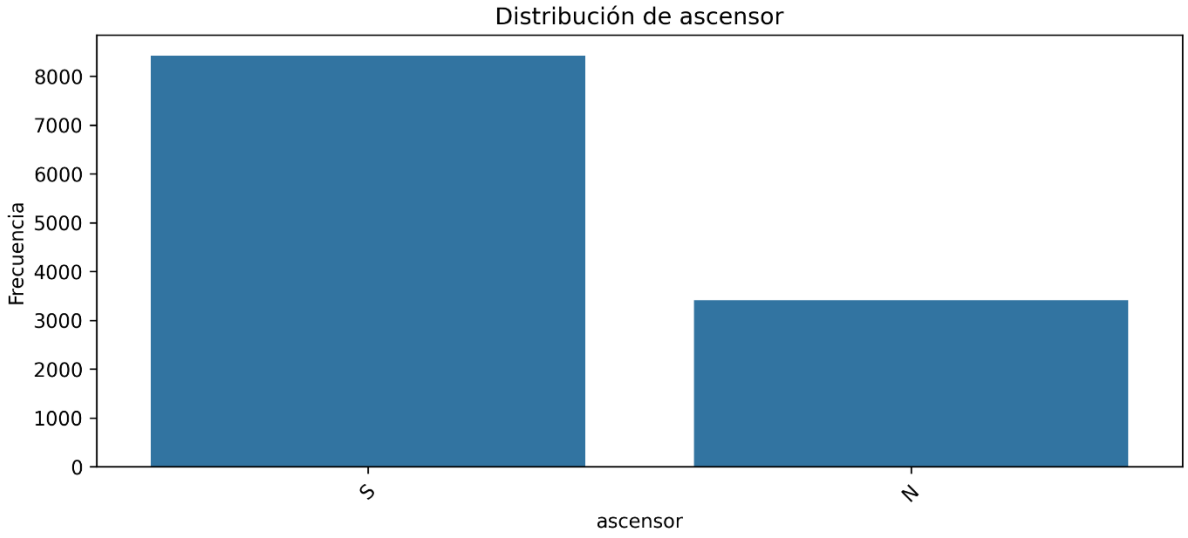
Este es un gráfico de barras que muestra la distribución de "localización" (tipo de ubicación).

- Eje X (horizontal): localización (EXTERIOR e INTERIOR).
- Eje Y (vertical): Frecuencia (número de propiedades en cada categoría).

La altura de cada barra representa cuántos anuncios entran en esa categoría.

Encontramos que la mayoría de los listados corresponden a propiedades exteriores, lo que indica que las propiedades interiores están infrarrepresentadas en la muestra.

5.3. Ascensor



Este es un gráfico de barras que muestra la distribución de los ascensores

- Eje X (ascensor):
  S = Sí (Sí, el edificio tiene ascensor) o N = No (Sin ascensor)
- Eje Y (Frecuencia): El número de observaciones (cuántos edificios/propiedades entran en cada categoría).

Este gráfico muestra que la mayoría de las propiedades del conjunto de datos tienen un elevador, mientras que una porción menor no.

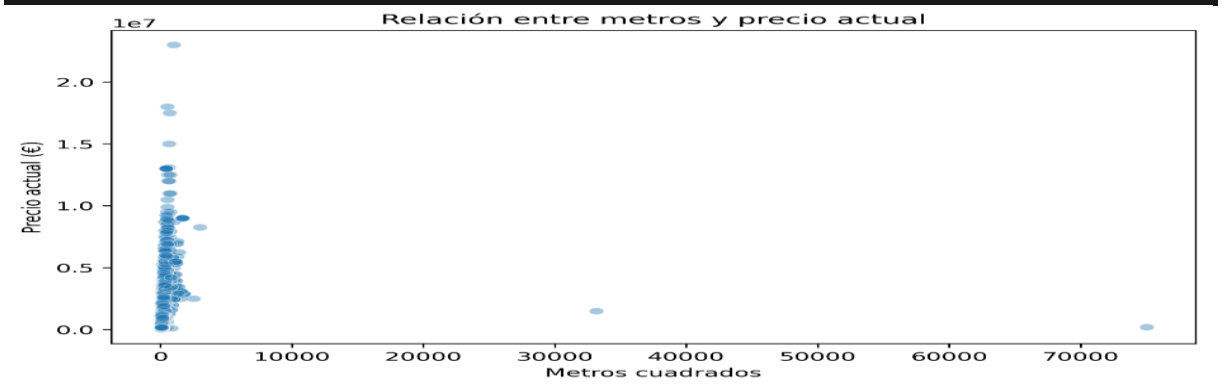
6. Análisis Bivariante

```python
# Importamos las librerías necesarias para análisis y visualización
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import os

# Cargamos el dataset limpio para estudiar relaciones entre variables
df = pd.read_csv("../data/Datos_clean.csv")

# Mostramos las primeras filas para comprobar que se ha cargado correctamente
df.head()
```

Relaciones entre variables del mercado inmobiliario de Madrid

6.1. Precios_cuadrados Análisis

```python
def precio_metros_cuadrados():
    # Analizamos cómo cambia el precio total según los metros cuadrados del inmueble
    # scatterplot permite ver la tendencia de crecimiento del precio con el tamaño
    plt.figure(figsize=(8,5))
    sns.scatterplot(data=df, x="metros", y="PrecioActual", alpha=0.4)
    plt.title("Relación entre metros y precio actual")
    plt.xlabel("Metros cuadrados")
    plt.ylabel("Precio actual (€)")

    # Guardado de la imagen
    plt.savefig("src/img/scatter_metros_precio.png", dpi=300, bbox_inches="tight")
    plt.show()
    plt.close()
```



Este es un diagrama de dispersión que muestra la relación entre el tamaño (metros cuadrados) y el precio actual (€).

- Eje X (horizontal): Metros cuadrados → el tamaño de la propiedad en metros cuadrados.
- Eje Y (vertical): Precio real (€) → el precio actual en euros.

Cada punto representa una propiedad.

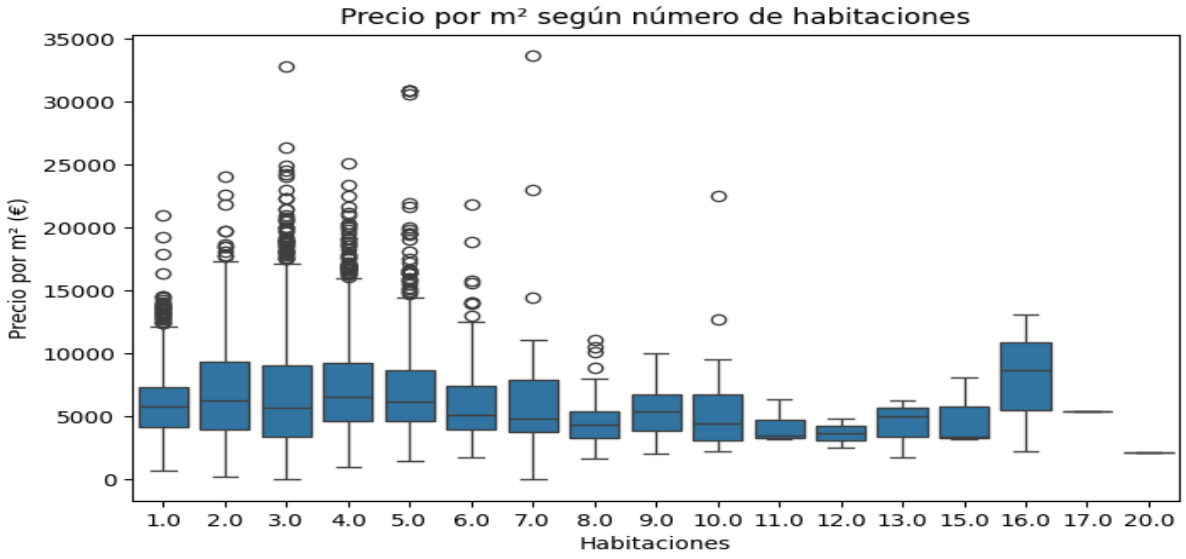
- Cuanto más a la derecha significa un punto, más grande es la propiedad.
- Cuanto más alto esté un punto, más cara es la propiedad.
- Un grupo de puntos significa muchas propiedades con tamaño y precio similares.

Según el análisis anterior, entendemos que hay una alta dispersión de precios para tamaños similares.

- La ubicación, el tipo de propiedad y las características probablemente importan más que el tamaño por sí solas.
- El conjunto de datos contiene valores atípicos (tipos especiales de terrenos (rurales, industriales), errores de datos, propiedades en ubicaciones o categorías muy diferentes) que merecen especial atención.

6.2. Habitaciones _ precio Análisis

```python
def habitaciones_precio():
    # Analizamos si el número de habitaciones influye en el precio por metro cuadrado.
    # boxplot muestra la variación del precio por m² según número de habitaciones
    plt.figure(figsize=(8,5))
    sns.boxplot(data=df, x="habitaciones", y="precio_m2")
    plt.title("Precio por m² según número de habitaciones")
    plt.xlabel("Habitaciones")
    plt.ylabel("Precio por m² (€)")

    # Guardado de la imagen
    plt.savefig("src/img/box_habitaciones_precio_m2.png", dpi=300, bbox_inches="tight")
    plt.show()
    plt.close()
```



Esta cifra es un diagrama de caja que muestra el precio por metro cuadrado (€/m^2) dependiendo del número de habitaciones (habitaciones).
Para cada número de habitaciones:

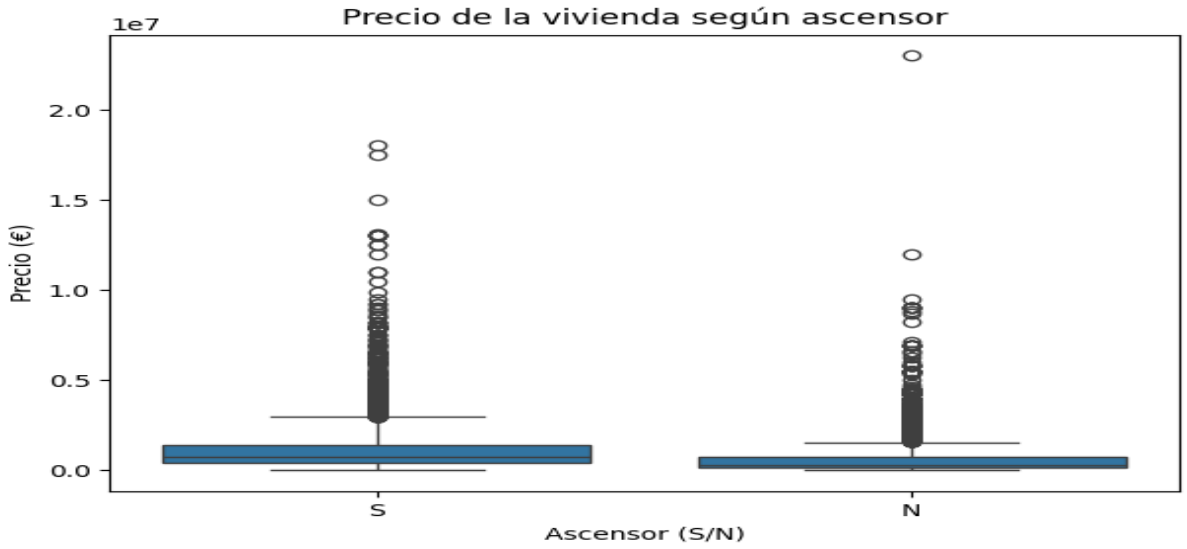
- Línea intermedia dentro de la caja → precio medio por m^2 (el valor "típico").
- Al final de la caja → percentil 25 (Q1).
- La parte superior de la caja → percentil 75 (Q3).
- Altura de la caja → Distribución del 50% medio de los precios (variabilidad).

Podemos concluir que:

- Los apartamentos más pequeños suelen costar más por m^2.
- A medida que aumenta el número de habitaciones, el precio por m^2 suele disminuir, con excepciones para propiedades de lujo.
- Los valores atípicos demuestran que la ubicación y la calidad importan mucho, no solo el tamaño

6.3. Ascensor_Precio Análisis

```python
def ascensor_precio():
    # Las viviendas con ascensor suelen tener mayor valor, especialmente si están en pisos altos.
    # boxplot muestra si hay diferencia de precio entre propiedades con o sin ascensor
    plt.figure(figsize=(7,5))
    sns.boxplot(data=df, x="ascensor", y="PrecioActual")
    plt.title("Precio de la vivienda según ascensor")
    plt.xlabel("Ascensor (S/N)")
    plt.ylabel("Precio (€)")

    # Guardado
    plt.savefig("src/img/box_ascensor_precio.png", dpi=300, bbox_inches="tight")
    plt.show()
    plt.close()
```



Este es un diagrama de caja que compara los precios de la vivienda en función de si el edificio tiene ascensor.

- Eje X (Ascensor S/N)
  S = Sí (con ascensor) , N = No (sin ascensor)
- Eje Y (Precio €)
  (1e7 significa que los valores se muestran en decenas de millones).

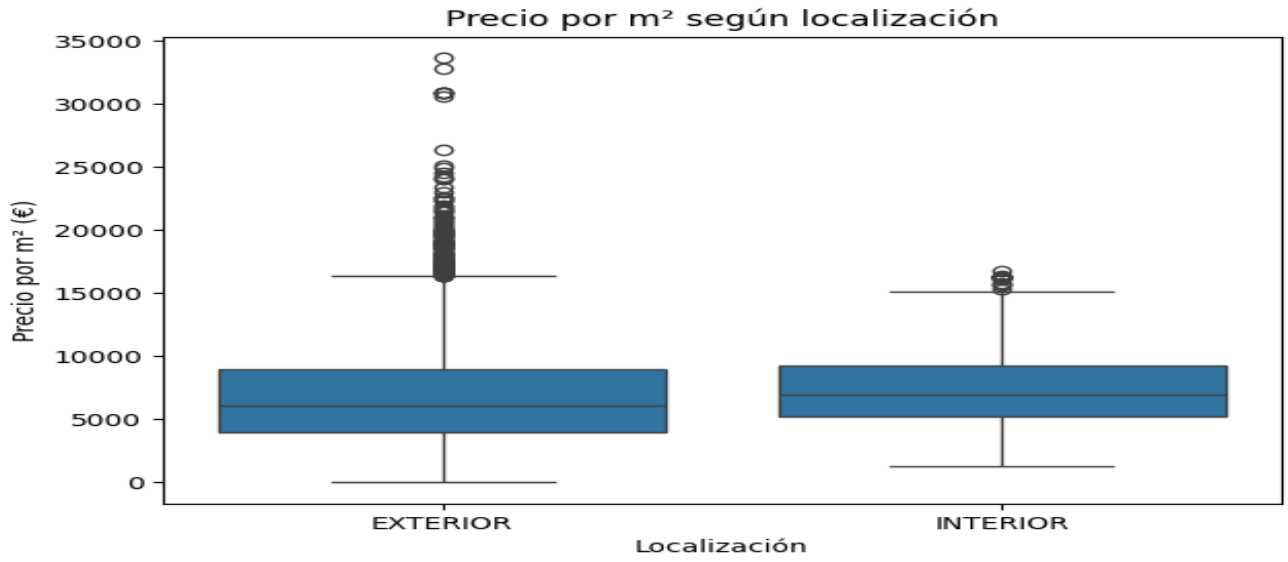
Para cada categoría (S o N):

- Línea media dentro del recuadro → Precio medio (el 50 % de las viviendas son más baratas, el 50 % más caras)
- Recuadro → 50 % medio de los precios (del percentil 25 al 75)
- Bigotes → Precios mínimos y máximos típicos (excluyendo los extremos)
- Puntos → Valores atípicos (propiedades inusualmente caras)

Podemos concluir que las propiedades con ascensor suelen ser más caras y sus precios varían más que las que no lo tienen.

6.4. Localización_precio Análisis

```python
def localizacion_precio():
    # Analizamos la relación entre la localización del inmueble y su precio por m².
    # Viviendas exteriores generalmente se venden más caras que las interiores
    plt.figure(figsize=(7,5))
    sns.boxplot(data=df, x="localizacion", y="precio_m2")
    plt.title("Precio por m² según localización")
    plt.xlabel("Localización")
    plt.ylabel("Precio por m² (€)")

    # Guardado
    plt.savefig("src/img/box_localizacion_precio_m2.png", dpi=300, bbox_inches="tight")
    plt.show()
    plt.close()
```



Este es un diagrama de caja, pero ahora compara el precio por metro cuadrado (€ / m^2) según el tipo de ubicación.

- Eje X (Localización)
  EXTERIOR: apartamentos que dan a la calle / exterior.
  INTERIOR: apartamentos que dan a patios interiores o al interior del edificio.
- Eje Y (Precio por m^2 €)

Para cada grupo (EXTERIOR, INTERIOR):

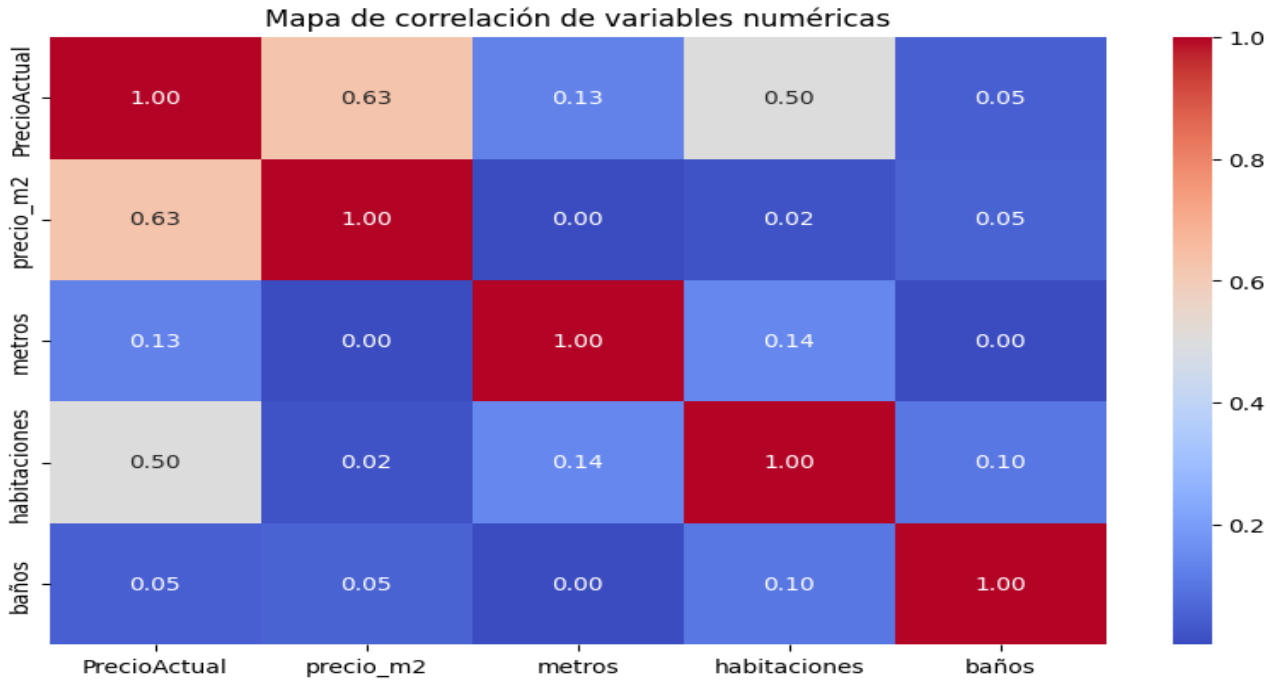
- Línea dentro de la caja → Precio medio por m^2.
- Caja → 50 % medio de los precios (percentil 25-75).
- Bigotes → Valores mínimos y máximos típicos.
- Puntos → Valores atípicos (muy caros €/m^2 en comparación con la mayoría de las viviendas)

Interpretación:

- Los apartamentos interiores tienden a ser más caros por m^2 de media.
- Los apartamentos exteriores muestran mucha más variabilidad, con casos tanto muy baratos como extremadamente caros.

7. Análisis Multivariante

```python
def mapa_calor():
    # Usamos un mapa de calor para ver qué variables tienen más relación entre sí.
    # La correlación más fuerte indica qué variables afectan más al precio
    plt.figure(figsize=(10,6))
    sns.heatmap(df[["PrecioActual", "precio_m2", "metros", "habitaciones", "baños"]].corr(),
                annot=True, cmap="coolwarm", fmt=".2f")
    plt.title("Mapa de correlación de variables numéricas")

    # Guardado
    plt.savefig("src/img/heatmap_correlaciones.png", dpi=300, bbox_inches="tight")
    plt.show()
    plt.close()
```



Este es un mapa de calor de correlación. Muestra la fuerza con la que las variables numéricas están relacionadas entre sí.

- Las filas y columnas son las mismas variables: Precio Actual, precio_m2, metros, habitaciones, baños.
- Cada celda muestra la correlación entre dos variables.
- Color: Rojo → correlación más fuerte / Azul → correlación más débil.
- La diagonal (1,00) siempre es roja porque cada variable está perfectamente correlacionada consigo misma.

Podemos concluir que:

- El precio depende más de factores relacionados con la calidad (precio por m^2, dormitorios) que del tamaño bruto.
- Los metros, las habitaciones y los baños no están fuertemente correlacionados, lo que sugiere una gran diversidad de distribuciones.

8. Conclusiones del análisis univariante (variables categóricas)

- **Zona**: La mayor concentración de anuncios se encuentra en los barrios de **Salamanca**, **Chamartín** y **Chamberí,** zonas de alto poder adquisitivo. Esto puede sesgar el dataset hacia precios elevados.
- **Localización**: La mayoría de las viviendas son **exteriores**, lo cual es un factor positivo que suele incrementar el precio.
- **Ascensor**: La presencia de ascensor es mayoritaria, pero todavía existe una parte relevante de viviendas sin ascensor, lo cual puede afectar al precio de pisos en plantas altas.

Estas observaciones sugieren que **ubicación** y **características del edificio** podrían tener un impacto significativo en el precio de la vivienda, lo que será analizado en la siguiente fase del proyecto.

9. Conclusiones del análisis bivariante

- **Superficie vs precio**: existe una relación positiva; a mayor tamaño, mayor precio, aunque se observan valores atípicos que podrían corresponder a edificios o errores.
- **Habitaciones**: el precio por m^2 no aumenta linealmente con las habitaciones, lo que indica que otros factores influyen más en el valor.
- **Ascensor:** las viviendas con ascensor presentan precios significativamente mayores, lo que confirma su influencia en el mercado.
- **Localización (interior/exterior):** las viviendas exteriores son más caras por m^2, lo que coincide con las preferencias del mercado.
- **Correlación general:** la variable que mayor influencia tiene sobre el precio total es el número de metros cuadrados.

Este análisis confirma que **tamaño** y **características del edificio** son factores clave en el precio de la vivienda en Madrid.

Esto nos permitirá detectar:

- Sesgos en el mercado inmobiliario
- Valores extremos (outliers)
- Concentración de valores según el tipo de vivienda

10. Visualización