Idealista nació en el año 2000, en un momento en el que el acceso a la información inmobiliaria en España era limitado y fragmentado. Sus fundadores vislumbraron el potencial de internet para transformar la forma en que las personas buscaban y compraban viviendas. Con esta visión en mente, crearon una plataforma online que reunía en un solo lugar miles de anuncios de inmuebles, ofreciendo a los usuarios una herramienta práctica y accesible para encontrar su hogar ideal.

En la actualidad, Idealista ofrece a través de Internet entre otros los servicios de portal inmobiliario en España, Andorra, Italia y Portugal.

El dataset contiene 16 columnas, cada una representando una variable de las propiedades, tales como: Cantidad de habitaciones, Distancia al centro, Cantidad de baños, si tiene piscina o no, si tiene ascensor, si tiene patio, etc etc etc.

En esta investigación se analizarán las siguientes hipótesis:

1) ¿Cómo influye la ubicación (barrio) en el precio de las propiedades?

Hipótesis: Propiedades en barrios céntricos o de alta demanda tienen precios significativamente más altos que en barrios, periféricos.

2. ¿Las amenidades como piscina, jardín y aire acondicionado están asociadas con precios más altos y un mayor número de baños?

Hipótesis: Las propiedades con piscina, jardín y aire acondicionado tienen precios más altos, especialmente si tienen más baños.

3. ¿Las propiedades con orientación sur y más habitaciones son más caras?

Hipótesis: Orientación de la propiedad, número de habitaciones y precio

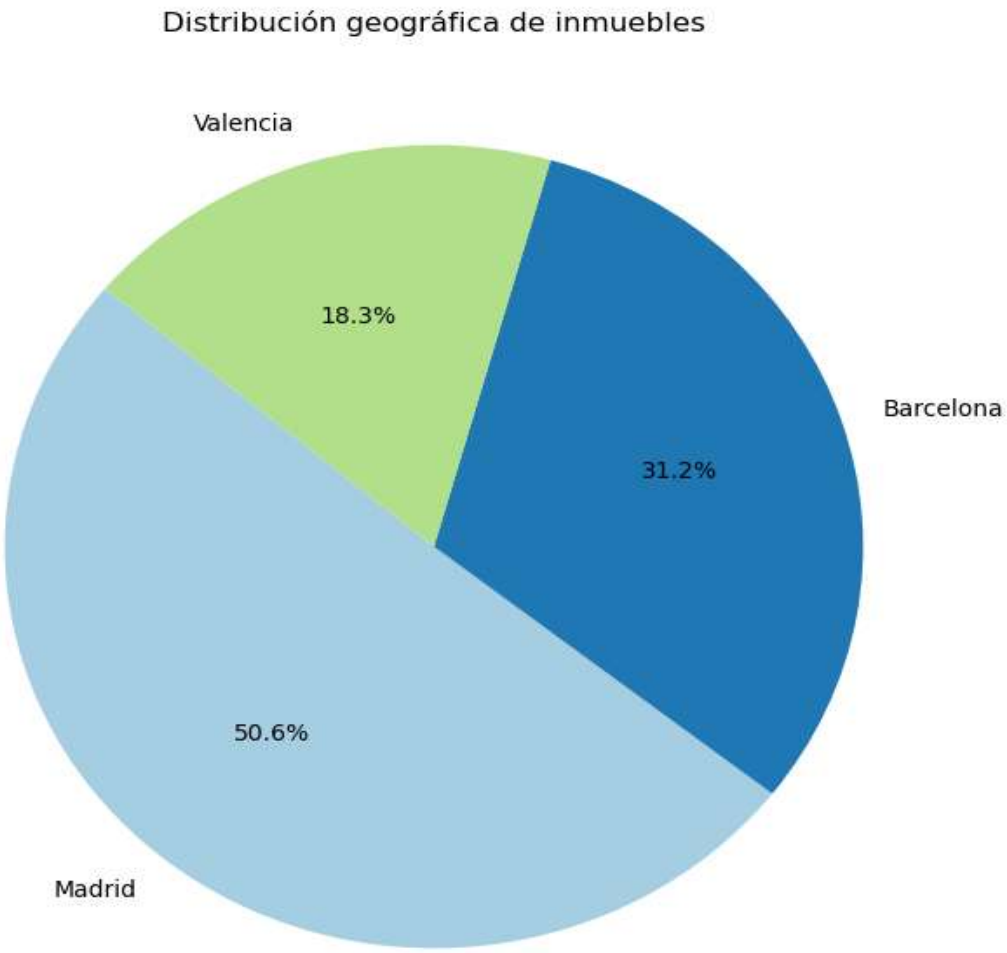
4. ¿Las propiedades con más habitaciones o baños tienen precios más altos?

Hipótesis: Propiedades con más habitaciones o baños tienden a ser más caras, pero el aumento en precio no es proporcional debido a otros factores (como el estado o la antigüedad).

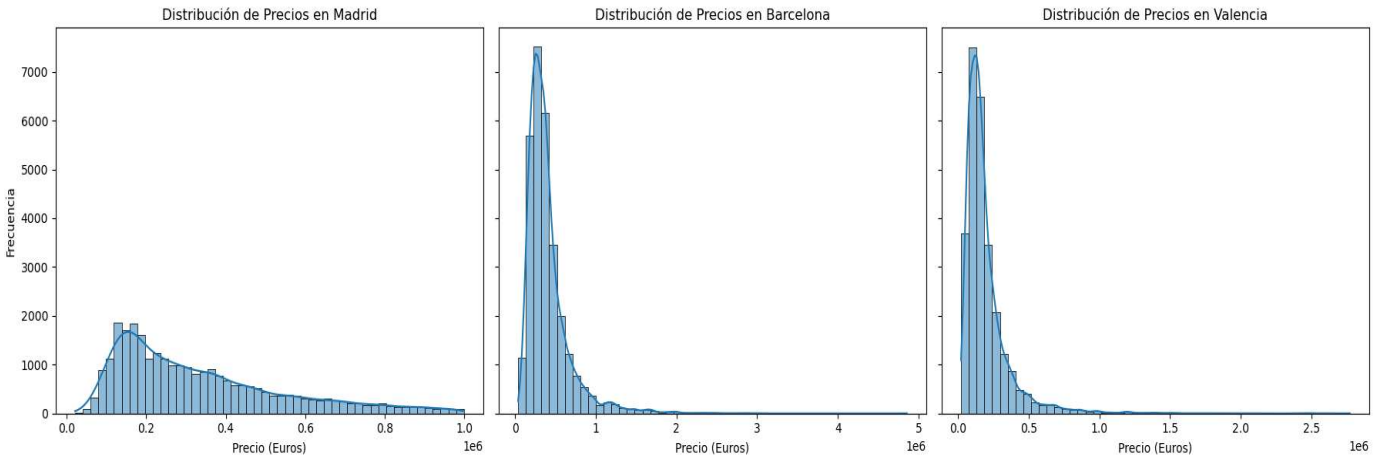
5. En qué ciudad los precios son más altos? Madrid, Barcelona o Valencia?

Hipótesis: Influye en el precio el tamaño de la ciudad y la capacidad habitacional

Útil para el tipo de inquilino, si son familia o personas solas o parejas.

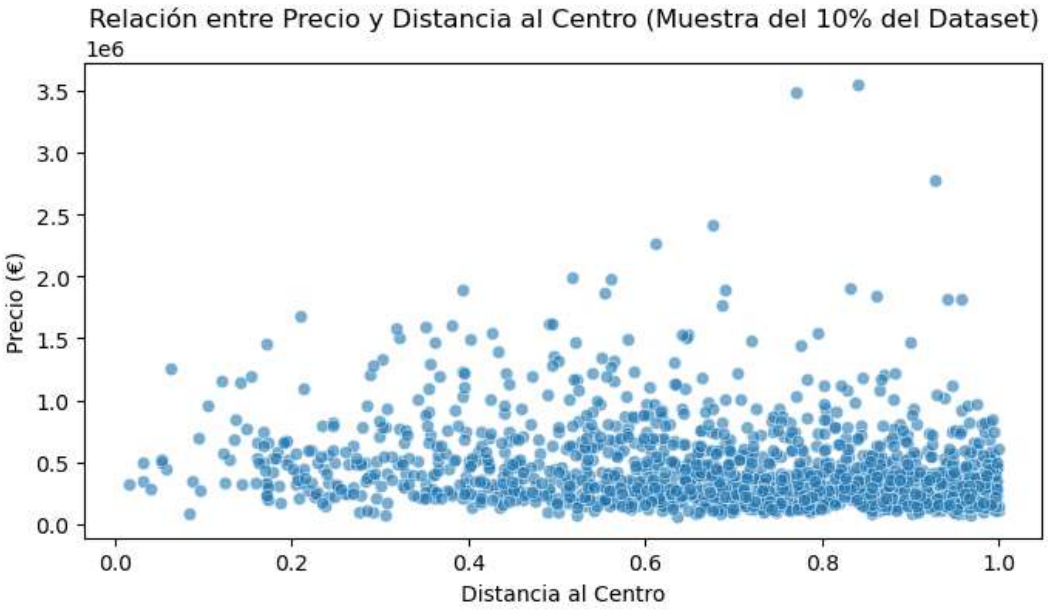


Se muestra la distribución de los precios en las tres ciudades bajo análisis.



¿Cómo influye la ubicación (barrio) en el precio de las propiedades?

Hipótesis: Propiedades en barrios céntricos o de alta demanda tienen precios significativamente más altos que en barrios periféricos.

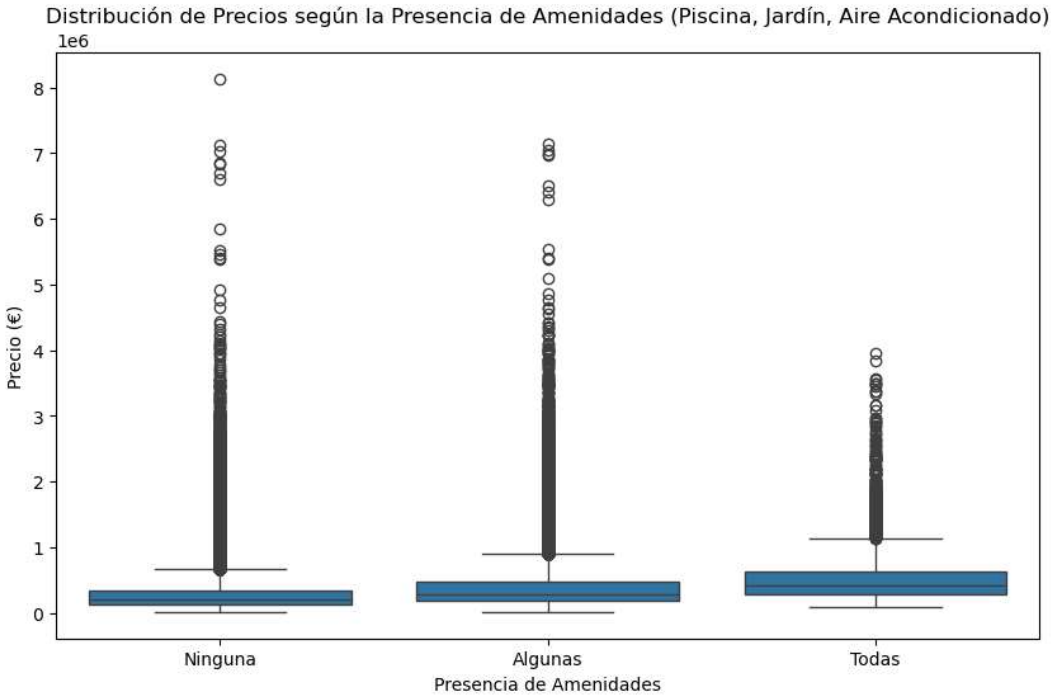


Se observa que en este dataset no se cumple la hipótesis de la cercanía de la propiedad al centro. Esto podría indicar que otros factores (como amenidades, número de habitaciones, o calidad catastral) influyen más en el precio.

¿Las amenidades como piscina, jardín y aire acondicionado están asociadas con precios más altos y un mayor número de baños?

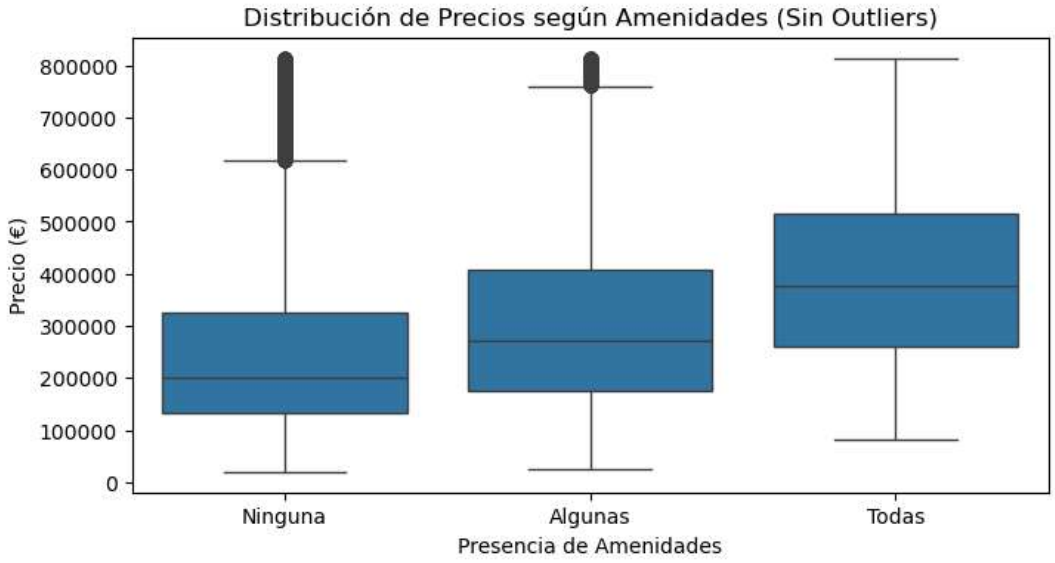
Hipótesis: Las propiedades con piscina, jardín y aire acondicionado tienen precios más altos.

Primero se creó una nueva variable categórica que indique si una propiedad tiene todas, algunas o ninguna de estas amenidades.



Se observa que la mediana de "Todas" es mayor que la de "Algunas" o "Ninguna" amenidad, lo cual significa que evidentemente encarecen el precio de las viviendas.

Se trata de eliminar outliers de la variable PRICE usando la técnica de los cuartiles (también conocida como el método del rango intercuartílico o IQR)



El nuevo boxplot no muestra puntos extremos (outliers) en la variable PRICE, lo que hace que las cajas y los bigotes reflejen mejor la distribución central de los datos.

Se mantiene la demostración de la Hipótesis, es decir, la mediana de precios en la categoría "Todas" (propiedades con piscina, jardín y aire acondicionado) sigue siendo más alta que en "Algunas" y "Ninguna", por lo tanto, las amenidades aumentan el precio.

La eliminación de outliers evita que valores extremos distorsionen las medianas o los rangos.

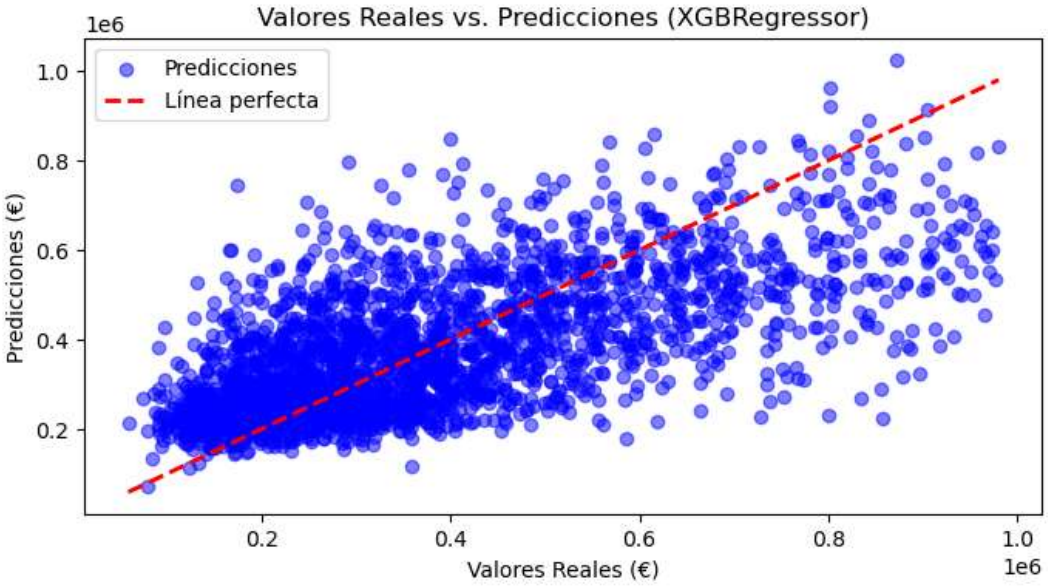
Comparado con el gráfico anterior (con outliers), las cajas podrían ser más compactas, y las diferencias entre categorías podrían ser más claras.

Regresión para Predecir Precios de Propiedades

En este análisis, se exploraron múltiples enfoques para predecir los precios de propiedades inmobiliarias (PRICE) utilizando un dataset que varió entre 13,773 y 149,923 observaciones. Inicialmente, un modelo RandomForestRegressor con un dataset de 149,923 filas, empleando variables como CONSTRROOMNUM, DISTANCE, BATHNUM, CADASTRALQUALITYID, HASPOOL, y HASGARDEN.

El modelo XGBRegressor demostró ser el más efectivo en el dataset más pequeño (13,773 filas), con un RMSE de 149,695.43 € (40.7% de la media), un MAPE de 36.41%, y un R^2 de 0.4303, superando al RandomForestRegressor con las mismas variables (RMSE de 164,301.98 €, MAPE de 39.44%, R^2 de 0.3137). Sin embargo, este desempeño es inferior al RandomForestRegressor inicial con un dataset más grande (RMSE de 123,699.89 €, MAPE de 43.00%, R^2 ~0.5-0.6), lo que sugiere que el tamaño del dataset influye significativamente en la precisión. La regresión lineal, con un RMSE de 336,479.65 € y un R^2 de 0.2277, fue el menos efectivo, destacando la necesidad de modelos no lineales como RandomForest y XGBoost para este problema. Ningún modelo alcanzó los estándares deseados para aplicaciones prácticas (MAPE < 20%, RMSE ~36,737-73,474 €, R^2 > 0.7-0.8), con errores promedio que representan entre el 40.7% y el 73.2% de la media de los precios.

Los modelos basados en árboles de decisión, como RandomForestRegressor y XGBRegressor, superaron a la regresión lineal, demostrando que las relaciones no lineales entre las variables predictoras (CONSTRROOMNUM, DISTANCE, BATHNUM, CADASTRALQUALITYID, HASSWIMMINGPOOL, HASGARDEN) y PRICE son significativas. XGBRegressor mostró una leve ventaja sobre RandomForestRegressor, gracias a su enfoque de boosting, que ajusta errores iterativamente.



Se usa un diagrama de dispersión (plt.scatter) para mostrar los valores reales (y_test) en el eje x y las predicciones (y_pred) en el eje y. La línea roja discontinua ('r--') representa la línea perfecta (donde los valores reales igualan las predicciones), sirviendo como referencia.

Dado que el RMSE es 149,695.43 € y el MAPE es 36.41%, se espera ver una dispersión amplia alrededor de la línea, especialmente para precios altos o bajos.

En conclusión, este análisis proporciona una base sólida para la predicción de precios inmobiliarios, destacando la superioridad de los modelos basados en árboles (especialmente XGBoost) y la importancia de un preprocesamiento adecuado y un dataset robusto. Aunque los resultados actuales no son lo suficientemente precisos para aplicaciones prácticas, ofrecen una dirección clara para futuros estudios que busquen desarrollar un modelo confiable y útil en el mercado inmobiliario.