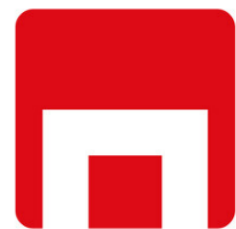




ANÁLISIS EXPLORATORIO DE DATOS (EDA)

Informe realizado por
Fran Rubio
Alba Rodríguez
Carlos D'Olhaberriague



The Bridge-DS

2025-2026

Índice

1. Introducción.....	4
1.1 ¿Por qué Airbnb?	4
1.2 Objetivo del estudio.....	4
1.3 Enfoque metodológico.....	4
2. Planteamiento del problema e hipótesis	5
3. Descripción de los datos.....	6
3.1 Origen de los datos.....	6
3.2 Estructura del dataset.....	6
3.3 Variables y tipos de datos	6
3.4 Variables clave.....	6
4. Preprocesado y limpieza de datos.....	6
4.1 Tratamiento de valores nulos.....	7
4.2 Detección y eliminación de duplicados	7
4.3 Conversión y ajuste de tipos de datos.....	7
4.4 Identificación y tratamiento de outliers	7
4.5 Dataset final para el análisis.....	7
5. Análisis Exploratorio de Datos (EDA).....	8
5.1 Análisis univariante.....	8
Distribución del precio.....	8
Tipo de alojamiento.....	9
Distribución por barrios.....	9
Cierre del análisis univariante	10
5.2 Análisis bivalente.....	10
Precio y barrio.....	10
Precio y tipo de alojamiento	11
Precio y capacidad del alojamiento	11
Precio y valoraciones de los usuarios	12
Precio y anfitrión verificado.....	13
Precio y número mínimo de noches	13
Cierre del análisis bivalente	14

5.3 Análisis multivariante.....	14
Correlación entre las principales variables de interés.....	14
Relación entre precio, capacidad y tipo de alojamiento.....	15
Relaciones multivariantes entre variables numéricas.....	15
6. Resultados y verificación de hipótesis	16
Hipótesis 1: El precio varía en función del barrio	16
Hipótesis 2: El tipo de alojamiento influye en el precio	16
Hipótesis 3: A mayor capacidad del alojamiento, mayor precio	16
Hipótesis 4: Una mejor valoración se asocia a un mayor precio	17
Hipótesis 5: El anfitrión verificado influye en el precio	17
Hipótesis 6: El número mínimo de noches afecta al precio.....	17
Conclusión general	17
7. Conclusiones.....	18
8. Recomendaciones y líneas futuras de análisis.....	18

1. Introducción

1.1 ¿Por qué Airbnb?

Airbnb se ha consolidado como una de las principales plataformas de alquiler turístico a nivel mundial, con una presencia especialmente relevante en grandes ciudades como Madrid. Se trata de un mercado dinámico y en constante evolución, caracterizado por una amplia diversidad de alojamientos, una elevada rotación de oferta y demanda y una gran cantidad de información disponible.

Este contexto convierte a Airbnb en un caso de estudio especialmente adecuado para un análisis exploratorio de datos, ya que permite trabajar con un volumen significativo de observaciones y múltiples variables relacionadas con el precio, la localización, las características de los alojamientos y el perfil de los anfitriones. En un mercado de estas características, el análisis de datos resulta clave para comprender su funcionamiento y facilitar la toma de decisiones informadas.

1.2 Objetivo del estudio

El objetivo principal de este trabajo es analizar y comparar las distintas características de los alojamientos ofertados en Airbnb en la ciudad de Madrid, con el fin de explorar cómo estas características pueden estar relacionadas con el precio por noche.

A través del análisis de variables como el barrio, el tipo de alojamiento, la capacidad, las valoraciones de los usuarios, la condición de anfitrión verificado y el número mínimo de noches requeridas, se pretende obtener una visión clara de cómo se estructura la oferta y qué factores pueden resultar relevantes para desenvolverse mejor dentro de este mercado.

1.3 Enfoque metodológico

El estudio se desarrolla mediante un Análisis Exploratorio de Datos (EDA), cuyo propósito es examinar, resumir y visualizar los datos para identificar patrones, relaciones entre variables y posibles comportamientos atípicos, así como contrastar hipótesis planteadas previamente.

Para ello, se ha utilizado un dataset público de Airbnb correspondiente a la ciudad de Madrid, que recoge información detallada sobre los alojamientos ofertados en la plataforma. Antes de realizar el análisis, los datos han sido sometidos a un proceso de preprocesado y limpieza con el objetivo de garantizar su coherencia y fiabilidad.

El análisis se estructura en distintas fases, incluyendo análisis univariante, bivariante y, cuando resulte pertinente, multivariante, apoyado en visualizaciones y estadísticos descriptivos. Los resultados obtenidos permitirán evaluar las hipótesis planteadas y extraer conclusiones basadas en los datos analizados.

2. Planteamiento del problema e hipótesis

El objetivo de este análisis exploratorio es examinar los factores que pueden estar asociados al precio de los alojamientos ofertados en Airbnb en la ciudad de Madrid, utilizando los datos disponibles en el dataset. El estudio se centra en analizar la posible influencia de distintas variables relacionadas con la localización, las características del alojamiento y el perfil del anfitrión, sin asumir relaciones previas y dejando que los datos guíen las conclusiones.

Para estructurar el análisis y orientar la exploración de los datos, se han definido las siguientes hipótesis de trabajo, que serán contrastadas en los apartados posteriores mediante análisis descriptivos y visualizaciones.

Hipótesis de trabajo:

Hipótesis 1: El precio de los alojamientos varía en función del barrio en el que se encuentran ubicados.

Hipótesis 2: El tipo de alojamiento ofertado (vivienda completa, habitación privada o compartida) influye en el precio por noche.

Hipótesis 3: Existe una relación entre la capacidad del alojamiento (número de huéspedes admitidos) y el precio.

Hipótesis 4: Las valoraciones de los usuarios podrían estar asociadas a diferencias en el precio de los alojamientos.

Hipótesis 5: La verificación del anfitrión (host verificado) podría influir en el precio del alojamiento.

Hipótesis 6: El número mínimo de noches requerido para la reserva podría afectar al precio por noche del alojamiento.

Estas hipótesis sirven como marco de referencia para el desarrollo del análisis exploratorio de datos y serán evaluadas de forma individual en el apartado de resultados y verificación de hipótesis, sin anticipar conclusiones en esta fase del trabajo.

3. Descripción de los datos

3.1 Origen de los datos

El análisis se ha realizado a partir de un dataset público de Airbnb correspondiente a la ciudad de Madrid, que recoge información sobre los alojamientos ofertados en la plataforma. El conjunto de datos incluye características relacionadas con el precio, la localización, el tipo de alojamiento, la capacidad, las valoraciones de los usuarios y las condiciones de reserva.

3.2 Estructura del dataset

El dataset está compuesto por registros individuales de alojamientos, donde cada fila representa un alojamiento ofertado en Airbnb en Madrid y cada columna corresponde a una variable descriptiva. El número de registros y variables disponibles permite realizar un análisis exploratorio adecuado y contrastar las hipótesis planteadas.

3.3 Variables y tipos de datos

Las variables incluidas pueden clasificarse en:

Variables numéricas, como el precio, la capacidad del alojamiento, las valoraciones o el número mínimo de noches.

Variables categóricas, como el barrio, el tipo de alojamiento o la condición de anfitrión verificado.

Una correcta identificación de los tipos de datos ha sido fundamental para el desarrollo del análisis posterior.

3.4 Variables clave

Dentro del conjunto de datos, el precio se considera la variable central del estudio, ya que actúa como referencia para analizar la influencia de otras características. Junto a esta, variables como el barrio, el tipo de alojamiento, la capacidad, las valoraciones, la verificación del anfitrión y el mínimo de noches requeridas resultan especialmente relevantes para el contraste de las hipótesis planteadas.

4. Preprocesado y limpieza de datos

Antes de realizar el análisis exploratorio, fue necesario llevar a cabo un proceso de preprocesado y limpieza de los datos con el objetivo de garantizar su coherencia, fiabilidad y adecuación al estudio. Esta fase es fundamental en un EDA, ya que la calidad de los resultados depende directamente de la calidad de los datos analizados.

4.1 Tratamiento de valores nulos

Se identificaron valores nulos en distintas variables del dataset. Para cada caso, se analizó su frecuencia y relevancia dentro del conjunto de datos. En función de la variable afectada y del volumen de valores ausentes, se optó por:

Eliminar registros con información incompleta cuando su impacto sobre el conjunto era reducido.

Mantener los valores nulos en aquellas variables donde su eliminación no resultaba adecuada, evitando introducir sesgos artificiales en los datos.

Las decisiones tomadas se realizaron siempre atendiendo a la naturaleza de la variable y a su importancia para el análisis posterior.

4.2 Detección y eliminación de duplicados

Se comprobó la existencia de posibles registros duplicados en el dataset. Aquellos registros que representaban el mismo alojamiento y no aportaban información adicional fueron eliminados, con el fin de evitar distorsiones en los análisis descriptivos y comparativos.

4.3 Conversión y ajuste de tipos de datos

Algunas variables requirieron la conversión de su tipo de dato para asegurar su correcta interpretación y análisis. Este proceso incluyó la transformación de variables numéricas y categóricas al formato adecuado, facilitando la aplicación de estadísticos y visualizaciones en las fases posteriores.

4.4 Identificación y tratamiento de outliers

Se analizaron posibles valores atípicos, especialmente en variables numéricas como el precio. Estos valores fueron identificados mediante análisis descriptivos y visualizaciones. En lugar de eliminarlos de forma automática, se evaluó su coherencia dentro del contexto del mercado, manteniéndolos cuando representaban situaciones plausibles y relevantes para el estudio.

4.5 Dataset final para el análisis

Tras completar el proceso de preprocesado y limpieza, se obtuvo un dataset depurado y coherente, adecuado para la realización del análisis exploratorio de datos. Este conjunto de datos final sirve como base para los análisis univariante, bivariante y multivariante desarrollados en los apartados siguientes.

5. Análisis Exploratorio de Datos (EDA)

El análisis exploratorio de datos constituye el núcleo del presente trabajo. En este apartado se estudian las principales variables del dataset mediante análisis univariante, bivalente y multivariante, con el objetivo de comprender su comportamiento individual y conjunto, así como de contrastar las hipótesis planteadas previamente.

5.1 Análisis univariante

El análisis univariante tiene como objetivo estudiar el comportamiento individual de cada variable del dataset, sin considerar su relación con otras variables. Este tipo de análisis permite comprender la distribución de los datos, identificar valores frecuentes, rangos habituales y posibles comportamientos atípicos, sirviendo como base para los análisis posteriores.

En este apartado se analizan algunas de las variables más representativas del conjunto de datos, centrándose en el precio, el tipo de alojamiento y la distribución geográfica de los anuncios.

Distribución del precio

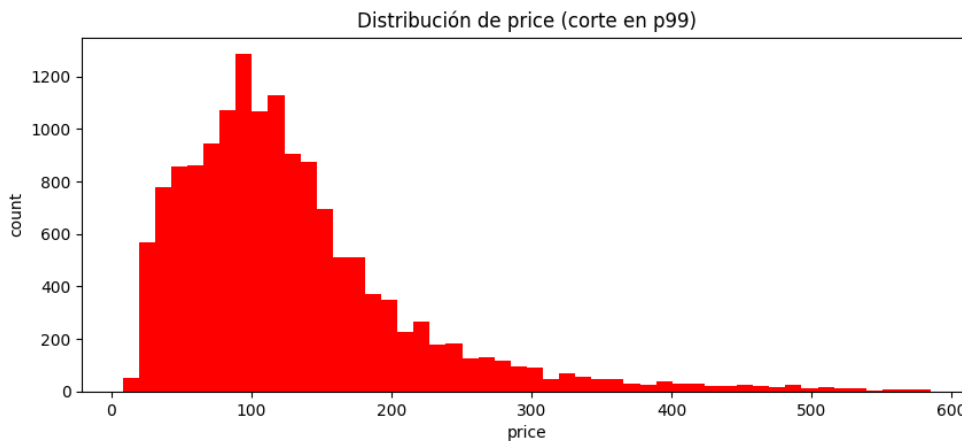


Ilustración 1 Análisis univariante del precio

La variable precio constituye uno de los elementos centrales del análisis, ya que será utilizada como referencia en gran parte del estudio. Para su análisis univariante se ha representado la distribución de los precios mediante un histograma, aplicando un corte en el percentil 99 con el fin de evitar que valores extremadamente altos distorsionen la visualización.

Esta representación permite observar la forma general de la distribución, identificar los rangos de precios más frecuentes y detectar la presencia de asimetría y valores atípicos dentro del conjunto de datos.

Tipo de alojamiento

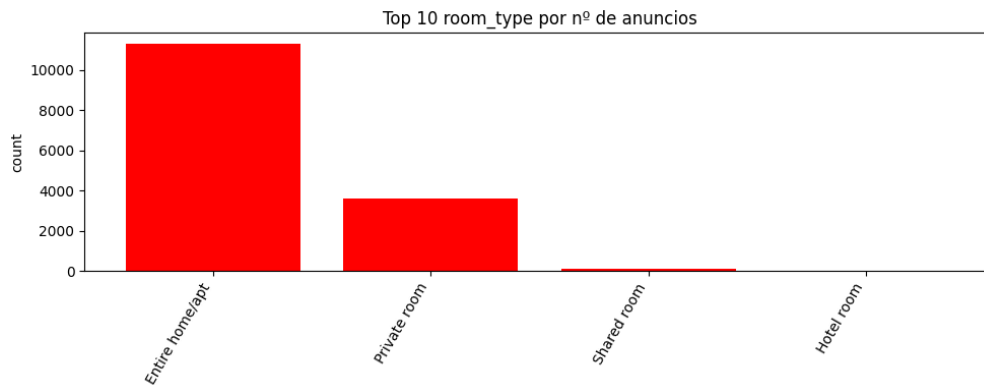


Ilustración 2 Análisis univariante sobre los tipos de habitación

A continuación, se analiza la variable correspondiente al tipo de alojamiento ofertado en la plataforma. Para ello, se ha utilizado un gráfico de barras que muestra el número de anuncios para cada categoría disponible en el dataset.

Este análisis permite conocer cómo se distribuye la oferta entre las distintas modalidades de alojamiento, proporcionando una visión general de las tipologías predominantes dentro del mercado de Airbnb en Madrid.

Distribución por barrios

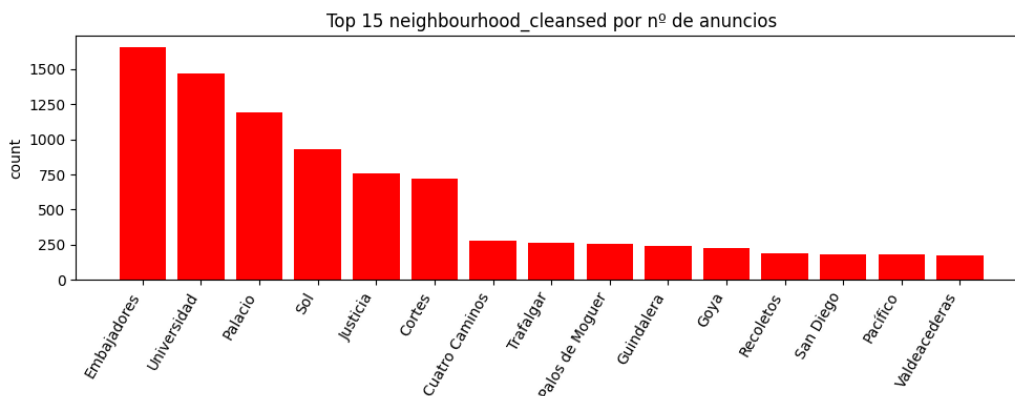


Ilustración 3 Análisis univariante de los barrios más populares por ocupación

Por último, se estudia la distribución geográfica de los alojamientos a través de la variable barrio. Dado el elevado número de categorías posibles, se han representado únicamente los 15 barrios con mayor número de anuncios, con el objetivo de facilitar la interpretación de los resultados.

El gráfico de barras resultante permite identificar las zonas de la ciudad con mayor concentración de alojamientos ofertados en Airbnb y ofrece una primera aproximación a la distribución espacial de la oferta dentro de Madrid.

Cierre del análisis univariante

El análisis univariante proporciona una visión general de las principales variables del dataset y permite contextualizar el comportamiento de los datos antes de abordar el estudio de relaciones entre variables. A partir de esta base, en el siguiente apartado se profundizará en el análisis bivariante, donde se analizará cómo interactúan estas variables entre sí, especialmente en relación con el precio.

5.2 Análisis bivariante

Precio y barrio

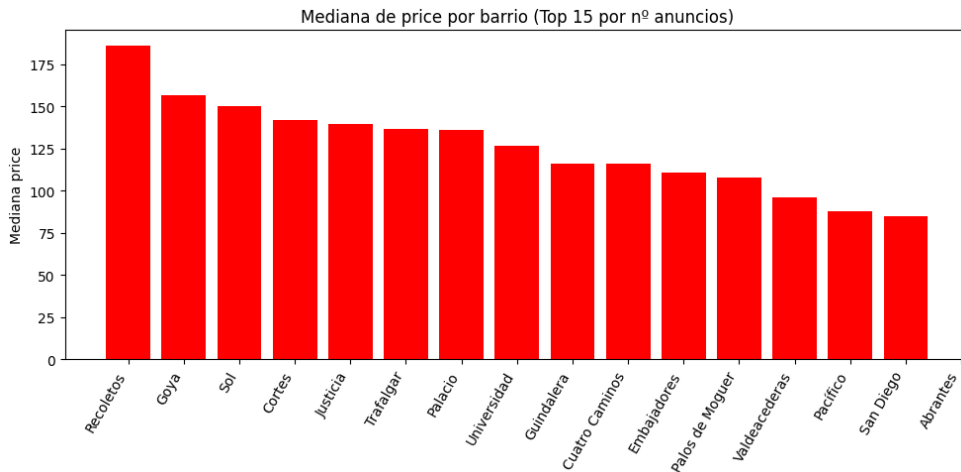


Ilustración 4 Análisis bivariante del precio por barrio

Para analizar la posible influencia de la localización en el precio, se estudia la relación entre el precio de los alojamientos y el barrio en el que se encuentran ubicados. Dado el elevado número de barrios presentes en el dataset, el análisis se ha limitado a aquellos con mayor número de anuncios, con el objetivo de facilitar la interpretación visual y evitar una sobrecarga de información.

Esta representación permite comparar la distribución de precios entre distintas zonas de la ciudad y observar posibles diferencias entre barrios.

Precio y tipo de alojamiento

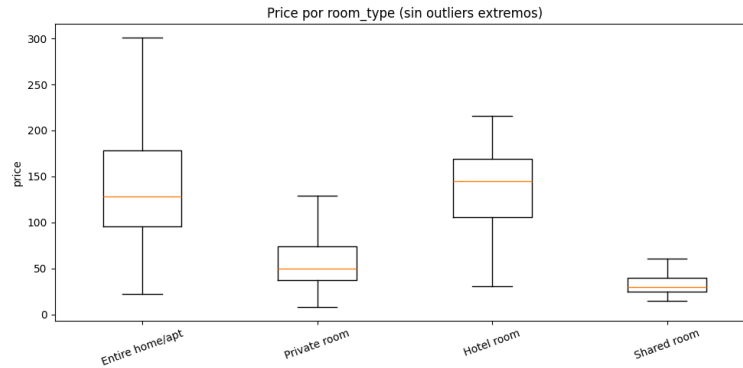


Ilustración 5 Análisis bivariante sobre el precio y el tipo de alojamiento

A continuación, se analiza la relación entre el precio y el tipo de alojamiento ofertado en la plataforma. Para ello, se comparan los precios entre las distintas categorías disponibles, lo que permite estudiar si existen diferencias asociadas a la modalidad del alojamiento.

Este análisis resulta especialmente relevante para comprender cómo se estructura la oferta en función del tipo de inmueble.

Precio y capacidad del alojamiento

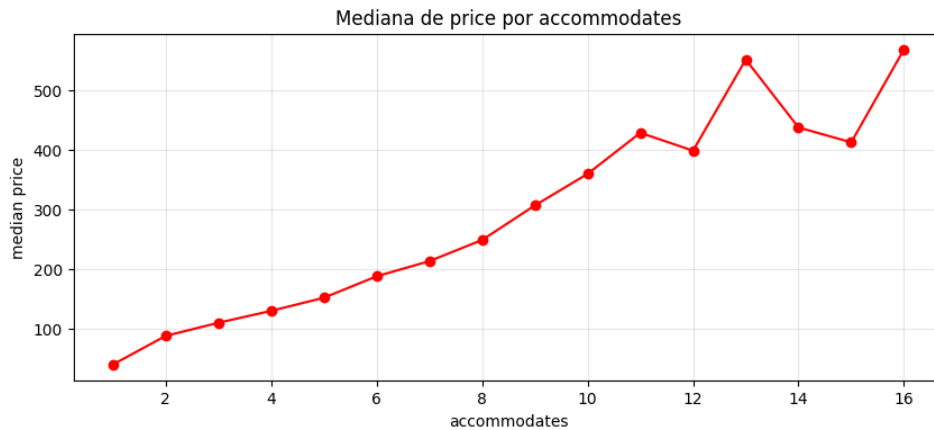


Ilustración 6 Análisis bivariante sobre el precio frente a la ocupación en huéspedes

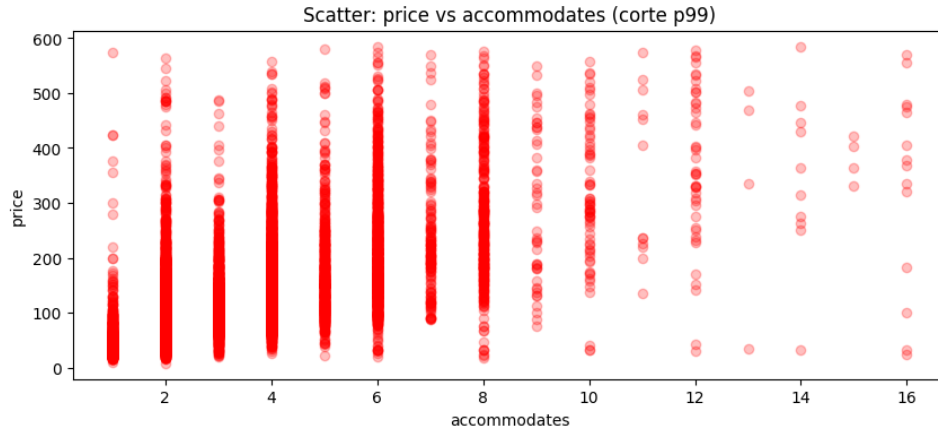


Ilustración 7 Volumen de cada registro para el precio frente ocupación en huéspedes

Se estudia la relación entre el precio y la capacidad máxima de huéspedes admitidos por cada alojamiento. Para ello, se utiliza una representación que permite observar cómo varía el precio a medida que aumenta la capacidad.

Este análisis permite explorar si existe una relación entre el tamaño funcional del alojamiento y su precio, se incluye además una distribución de los datos para cada valor apreciándose así el volumen de registros para cada marca.

Precio y valoraciones de los usuarios

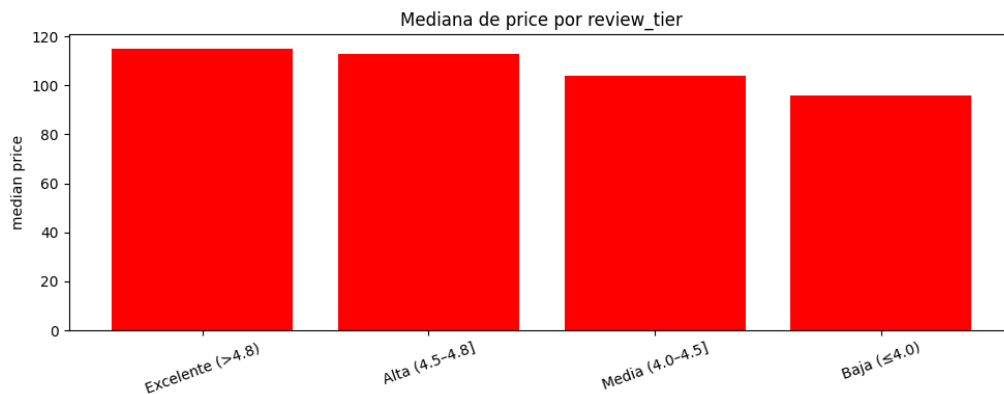


Ilustración 8 Relación entre el precio y las valoraciones del apartamento

En este apartado se analiza la relación entre el precio de los alojamientos y las valoraciones otorgadas por los usuarios. El objetivo es observar si se aprecian diferencias en el precio asociadas a distintos niveles de valoración.

Esta comparación permite explorar el posible papel de la percepción del usuario en la fijación de precios.

Precio y anfitrión verificado

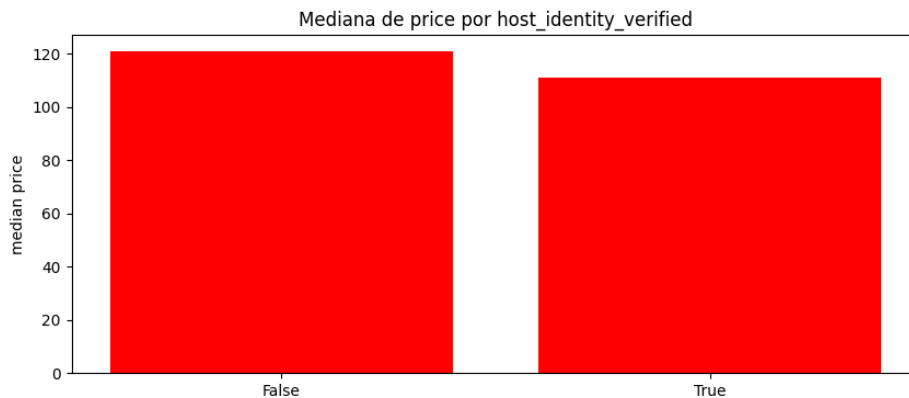


Ilustración 9 Análisis bivalente entre el precio y la verificación del propietario

Se analiza la relación entre el precio y la condición de anfitrión verificado, comparando los precios de los alojamientos en función de si el anfitrión cuenta o no con dicha verificación.

Este análisis permite estudiar si la verificación del anfitrión podría estar asociada a diferencias en el precio del alojamiento.

Precio y número mínimo de noches

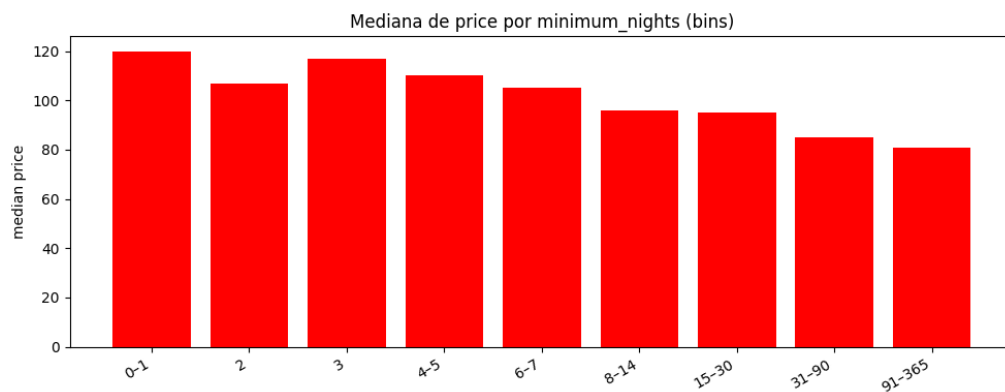


Ilustración 10 Análisis bivalente entre el precio y el mínimo número de noches de ocupación

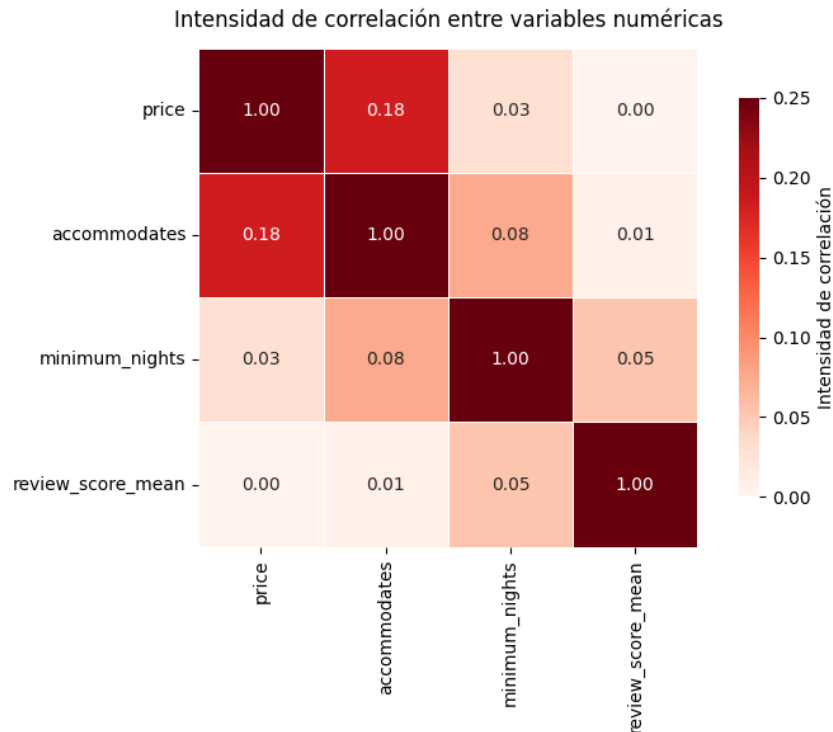
Por último, se analiza la relación entre el precio y el número mínimo de noches requerido para realizar una reserva. Esta comparación permite explorar si las condiciones de estancia mínima están relacionadas con el precio por noche ofertado.

Cierre del análisis bivalente

El análisis bivalente permite profundizar en las relaciones entre el precio y las principales variables del dataset, proporcionando una base sólida para el contraste de las hipótesis planteadas. Los resultados observados en este apartado serán evaluados de forma detallada en el siguiente apartado de resultados y verificación de hipótesis.

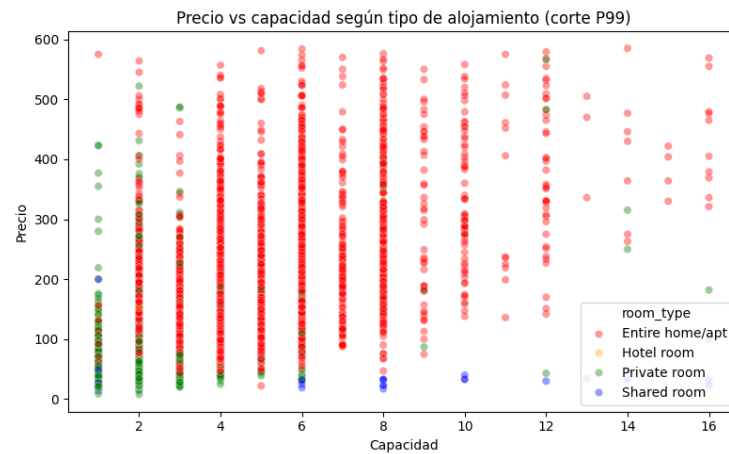
5.3 Análisis multivariante

Correlación entre las principales variables de interés



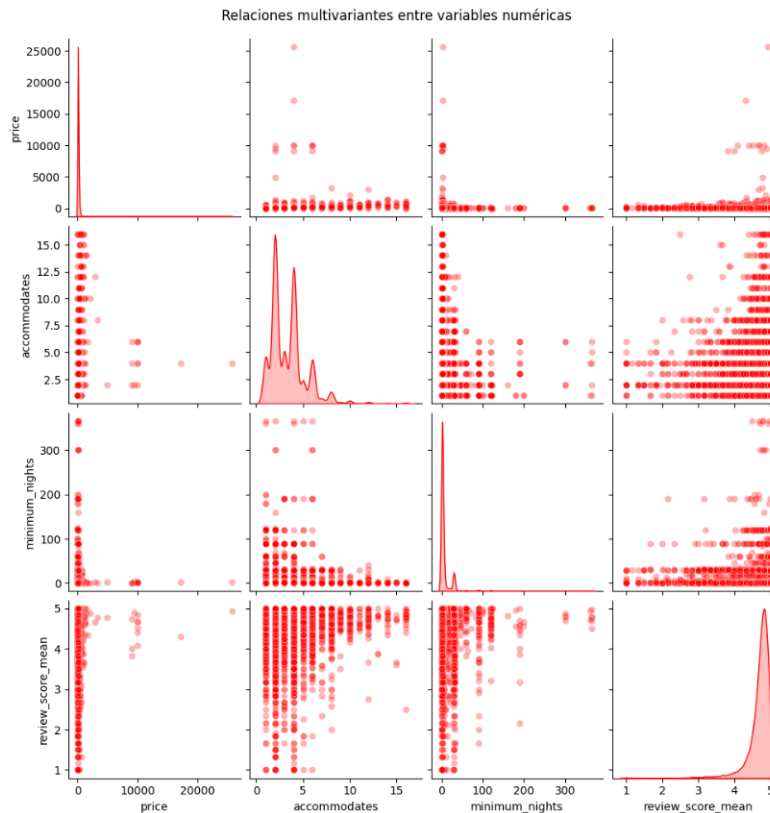
Este gráfico muestra la intensidad de la correlación entre las principales variables numéricas del estudio. Se observa que, en general, las correlaciones son bajas, lo que indica que el precio no depende de una única variable, sino de la combinación de varios factores. La relación más destacable se da entre el precio y la capacidad del alojamiento, aunque sigue siendo moderada.

Relación entre precio, capacidad y tipo de alojamiento



Este gráfico muestra cómo el precio tiende a aumentar con la capacidad del alojamiento, observándose diferencias claras entre los distintos tipos de alojamiento. El corte en el percentil 99 permite una mejor interpretación visual al reducir la influencia de valores extremos.

Relaciones multivariantes entre variables numéricas



La visualización conjunta de las variables numéricas no muestra relaciones lineales fuertes, reforzando la idea de que el precio está condicionado por la interacción de varios factores y no por una única variable.

6. Resultados y verificación de hipótesis

En este apartado se presentan los principales resultados obtenidos a partir del análisis exploratorio de datos y se procede a la verificación de las hipótesis planteadas inicialmente. Para cada hipótesis se resume brevemente el análisis realizado y se indica si los resultados observados permiten aceptarla o rechazarla, siempre desde una perspectiva descriptiva propia de un EDA.

Hipótesis 1: El precio varía en función del barrio

El análisis bivalente entre el precio y el barrio muestra diferencias claras en la distribución de precios entre las distintas zonas de la ciudad. Al comparar los barrios con mayor número de anuncios, se observa que existen barrios con precios medianos significativamente más altos que otros, lo que indica una influencia de la localización en el precio de los alojamientos.

Conclusión:

Los resultados apoyan la hipótesis de que el precio de los alojamientos varía en función del barrio, confirmando la relevancia de la localización dentro del mercado de Airbnb en Madrid.

Hipótesis 2: El tipo de alojamiento influye en el precio

El análisis del precio según el tipo de alojamiento revela diferencias evidentes entre las distintas modalidades ofertadas. Las distribuciones de precios muestran que no todas las categorías presentan el mismo comportamiento, existiendo variaciones claras entre viviendas completas, habitaciones privadas y otras tipologías.

Conclusión:

La hipótesis se considera verificada, ya que el tipo de alojamiento está asociado a diferencias en el precio por noche.

Hipótesis 3: A mayor capacidad del alojamiento, mayor precio

El análisis de la relación entre el precio y la capacidad del alojamiento muestra una tendencia general en la que los alojamientos con mayor capacidad presentan precios más elevados. Aunque la relación no es perfectamente lineal y existe dispersión, el patrón general apunta a una asociación positiva entre ambas variables.

Conclusión:

Los resultados respaldan la hipótesis de que la capacidad del alojamiento influye en el precio, observándose una relación positiva entre ambas variables.

Hipótesis 4: Una mejor valoración se asocia a un mayor precio

El análisis bivalente entre el precio y las valoraciones de los usuarios indica que los alojamientos con mejores valoraciones tienden a presentar precios más elevados. No obstante, la relación observada no es extremadamente marcada y existe solapamiento entre los distintos niveles de valoración.

Conclusión:

La hipótesis se considera parcialmente verificada, ya que se observa una asociación entre valoración y precio, aunque con una intensidad moderada.

Hipótesis 5: El anfitrión verificado influye en el precio

Al comparar los precios de los alojamientos en función de la condición de anfitrión verificado, apenas se observan diferencias entre ambos grupos. Los alojamientos no presentan una correlación directa con la variable

Conclusión:

Los resultados rechazan la hipótesis de que la verificación del anfitrión está asociada a diferencias en el precio del alojamiento.

Hipótesis 6: El número mínimo de noches afecta al precio

El análisis de la relación entre el precio y el número mínimo de noches requeridas muestra que los alojamientos con estancias mínimas más largas tienden a presentar precios por noche diferentes respecto a aquellos con menor restricción. Aunque existe dispersión, se identifican patrones que sugieren una relación entre ambas variables.

Conclusión:

La hipótesis se considera verificada, ya que el número mínimo de noches requerido parece estar relacionado con el precio por noche del alojamiento.

Conclusión general

A partir del análisis exploratorio realizado, se puede concluir que el precio de los alojamientos en Airbnb en Madrid está influido por múltiples factores, entre los que destacan la localización, el tipo de alojamiento, la capacidad, las valoraciones de los usuarios, la verificación del anfitrión y las condiciones de estancia mínima. Estos resultados confirman la utilidad del EDA como herramienta para comprender el comportamiento del mercado y proporcionan una base sólida para la formulación de recomendaciones, que se presentan en el apartado siguiente.

7. Conclusiones

El análisis exploratorio de datos realizado sobre los alojamientos ofertados en Airbnb en la ciudad de Madrid ha permitido comprender cómo se estructura la oferta y qué factores están asociados a las diferencias de precio entre los distintos anuncios. A través del estudio univariante, bivariante y multivariante, se ha obtenido una visión global del comportamiento de las principales variables del dataset.

Los resultados muestran que el precio de los alojamientos no depende de un único factor, sino que está influido por un conjunto de características relacionadas tanto con la localización como con el tipo de alojamiento y las condiciones de la reserva. Variables como el barrio, el tipo de alojamiento, la capacidad, las valoraciones de los usuarios, la verificación del anfitrión y el número mínimo de noches requeridas han demostrado ser relevantes para entender la variabilidad de precios en el mercado.

Este trabajo pone de manifiesto la utilidad del Análisis Exploratorio de Datos como herramienta para extraer información valiosa a partir de conjuntos de datos reales, permitiendo identificar patrones y relaciones sin necesidad de recurrir a modelos predictivos. Las conclusiones obtenidas proporcionan una base sólida para una mejor comprensión del mercado de Airbnb en Madrid y pueden servir de apoyo para la toma de decisiones fundamentadas en datos.

8. Recomendaciones y líneas futuras de análisis

Aunque el análisis exploratorio realizado permite obtener una visión clara del comportamiento del mercado de Airbnb en Madrid, es importante señalar algunas limitaciones del estudio y proponer posibles líneas de análisis futuras que podrían enriquecer los resultados obtenidos.

En primer lugar, el análisis no tiene en cuenta factores estacionales, como variaciones de la demanda en función de la época del año, eventos puntuales o periodos vacacionales, que podrían influir de manera significativa en los precios de los alojamientos. La incorporación de variables temporales permitiría profundizar en este aspecto y analizar posibles patrones estacionales.

Asimismo, el estudio se basa en una instantánea del mercado correspondiente al momento de recogida del dataset, sin considerar la evolución temporal de los precios ni los cambios en la oferta. Un análisis longitudinal permitiría evaluar tendencias a lo largo del tiempo y detectar cambios estructurales en el mercado.

Por otro lado, el enfoque del trabajo es exclusivamente descriptivo, propio de un Análisis Exploratorio de Datos. Como posible ampliación, los resultados obtenidos podrían servir como base para el desarrollo de modelos predictivos que estimen el precio de los alojamientos a partir de sus características, siempre tras una validación adecuada de los datos.

Finalmente, la incorporación de información externa, como indicadores turísticos, datos socioeconómicos por barrio o normativa local, podría aportar un mayor contexto al análisis y mejorar la interpretación de los resultados.