# Análisis de los Alquileres de Alojamientos Turísticos No Hosteleros en Málaga: Factores, Tendencias Y Excesos

## Objetivos del trabajo, marco y justificación

El objetivo de este trabajo es realizar un análisis del mercado de alquiler de alojamientos turísticos no hosteleros en la ciudad de Málaga que se anuncian en la plataforma Airbnb, identificando los factores determinantes del precio de los anuncios de estos alojamientos, identificar a su vez los barrios con mayor interés económico y el efecto que tiene a su vez la localización del alojamiento, todo esto con el objetivo final de crear un modelo que permita identificar si el precio de un alojamiento por noche anunciado es excesivo según el precio de mercado o no.

Que la provincia de Málaga es una de las más turísticas de nuestro país es algo de sobras conocido, y ello se extrapola a su capital, la ciudad de Málaga.

Durante todo el siglo XX, y especialmente a partir del aperturismo en el franquismo en los años 60 y el incentivo al crecimiento del turismo, la Costa del Sol empezó a tener como un pilar de su economía el sector turístico.

Durante el primer cuatrimestre la Costa Del Sol experimentó un incremento del 8,1% en la estancia de viajeros en establecimientos reglados y un 13,5% en los no reglados, es decir, en lo que se conoce como **alojamientos turísticos no hoteleros** (Habitaciones, casas, apartamentos de particulares, etc), y ¿Cuál es la plataforma referente en el anuncio y alquiler de este tipo de alojamientos? La plataforma Airbnb. Será sobre este sector sobre el que se desarrollará este trabajo.

El crecimiento mencionado no está exento de controversia. Entre otras cosas, la presencia de Airbnb en las ciudades provoca gentrificación, desplazamiento de la población local de los centros de las ciudades y de las ciudades mismas, así como crecimientos desmedidos de los precios del alquiler de larga estancia, pese a ser Airbnb una plataforma de alquileres de corta estancia al reducir la oferta tal y como muestran Wachsmuth y Weisler. Estos efectos han provocado diversas manifestaciones por parte de los malagueños por los precios.

No cabe duda de que los alojamientos turísticos no hoteleros son un tema de actualidad en Málaga.

## Objetivo y metodología

Desde un punto de vista económico, y viendo que los alojamientos turísticos no hoteleros son un tema de actualidad en Málaga cabe preguntarse ¿Cómo está dicho mercado?, ¿Cuáles son las características del mismo?, ¿Qué factores determinan los mismos? Con el objetivo de llegar a obtener una **Visión integral del mercado de alquiler de alojamientos turísticos no hoteleros en Málaga**

El objetivo de este trabajo es realizar un análisis del mercado de alquiler de alojamientos turísticos no hoteleros anunciados en Airbnb en la ciudad de Málaga, buscando identificar y comprender que características presentes en los anuncios de la plataforma resultan más interesantes y relevantes desde la perspectiva del posible cliente.

Se hará foco especialmente en los efectos sobre el precio y en el hallazgo de los determinantes del precio por noche de los alojamientos en base a las variables que se pueden observar desde la interfaz de los anuncios.

Y finalmente se desarrollará un modelo con estos datos que responda a la pregunta “¿Es el precio de un determinado anuncio adecuado o excesivo?”

Los pasos que se han seguido para hacer el análisis en este trabajo son los siguientes: **Análisis de la oferta => Análisis descriptivo => Modelo de regresión=>Análisis Clúster=>Modelo Logit**

## Análisis de la oferta

Para comenzar, se quiere responder a la pregunta *“¿Qué barrios son más atractivos turísticamente?*”.

Para ello, el procedimiento que se sigue en este trabajo será estudiar en que barrios hay más anuncios, ya que, es en aquellos barrios con más demanda de alojamientos donde habrá mayor oferta de los mismo.

En el gráfico aquí presente se han representado el número de anuncios en los barrios de Málaga de forma que a mayor número de anuncios, mayor tamaño y mayor tendencia al rojo tendrá la burbuja representada.

Destaca, sobre todos los demás, el barrio **Centro**, superando los 5000 anuncios frente a los demás que todos tienen menos de 1000. Hechos los cálculos y como es indicado en el trabajo escrito, **Centro** acapara el 66,67% de la oferta de anuncios presentes en Málaga.

Al estar este mercado en crecimiento, y este ser el barrio con mayor oferta, también es uno de los más demandados, por lo que los precios en él parece que tenderán a ser más altos que en el resto, por lo que parece intuirse que el barrio en que se encuentre el alojamiento turístico no hostelero anunciado será uno de los factores importantes en la determinación del precio.

Teniendo en cuenta la concentración significativa de anuncios en el barrio Centro y su impacto potencial en los precios, en el siguiente apartado se va a mostrar el análisis realizado al conjunto de datos enfocado en el precio, para ver de forma más detallada lo que ocurre (Wachsmuth & Weisler, 2018).

## Análisis Descriptivo

En este apartado, se va a mostrar el análisis realizado al conjunto de datos. La variable sobre la que se va a seguir el estudio en todo momento es el precio.

La mayor parte de los anuncios se acumulan en el intervalo [0, 250] euros por noche, habiendo un decrecimiento muy pronunciado a partir de ese precio, y si se observa el boxplot de los precios, es a partir de los 250 euros cuando empiezan los outliers.

Si se concreta este análisis barrio a barrio se observa lo siguiente:

Excepto en **Campanillas**, en todos los barrios hay outliers por encima de las franjas marcadas por el gráfico, y en ningún caso hay outliers por debajo. **Según Kimes y Wirtz (2015), la fijación de precios en el mercado de alojamientos tiende a presentar outliers cuando los precios superan significativamente el valor de mercado.** Los barrios que presentan mayor presencia de outliers son **Centro** y **Este**, por tanto, de acuerdo a Kimes y Wirtz, estos serán los barrios con el precio más superior al precio de mercado.

Si estos resultados se exponen en forma de distribución se tienen los gráficos siguientes:

* En el primer gráfico (Gráfico A), debido a la inmensa mayoría de anuncios en Centro, acapara toda la información y no es posible extraer mucho.
* En visos de una mejor apreciación, se ha extraído una muestra aleatoria de 200 alojamientos de los barrios que superan los 200 anuncios y se ha representado la distribución (Gráfico B). Aquí destaca Bailén-Miraflores.
* Si se extraen los barrios con más de 200 anuncios, en vez de escoger una muestra aleatoria de los mismos, se obtiene el tercer gráfico (Gráfico C), obteniéndose una distribución muy similar al primer tramo de la distribución mostrada al principio de este análisis, además, al eliminarse los barrios con más propiedades, también se eliminan los outliers más altos, es decir, que **Los barrios con mayor nº de anuncios son los que tienen mayores precios, lo que indica que la demanda se concentra en ellos y, pese a tener la mayor oferta, no es suficiente para cubrirla.**

## Modelo de Regresión

Yendo ahora a ver que determina el precio por noche de un alojamiento turístico no hotelero en Airbnb, a partir de las variables que se pueden apreciar en sus anuncios, se ha hecho un modelo de regresión con el objetivo de ver cuáles de estas variables son significativas.

Estos son los resultados del modelo, generado con el 80% de los datos y testeado con el 20% restante:

### Gráfico A: Residuals vs Fitted

* **Conclusión**:
  + El gráfico muestra la dispersión de los residuos en relación con los valores ajustados. La ligera curva en los residuos sugiere que hay patrones no capturados por el modelo lineal, lo que contribuye a la magnitud del error estándar residual.
  + La dispersión de los residuos alrededor de la línea horizontal indica que una gran parte de la variabilidad en el precio (60.54%) no es explicada por el modelo.
  + Los patrones en los residuos sugieren que hay aspectos de la variabilidad en los datos que no están siendo capturados adecuadamente, lo que justifica un R2R^2R2 ajustado ligeramente menor.

### Gráfico B: Normal Q-Q

* **Conclusión**:
  + El gráfico muestra cómo se distribuyen los residuos en comparación con una distribución normal. Las desviaciones en los extremos sugieren que el modelo no captura completamente todas las características de los datos, lo cual es consistente con un R2R^2R2 que no es extremadamente alto.

### Gráfico C: Scale-Location

* **Conclusión**:
  + La tendencia ascendente en este gráfico sugiere heterocedasticidad, es decir, que la variabilidad de los residuos aumenta con los valores ajustados. Esto implica que el error estándar residual podría no ser constante a lo largo de todos los valores de la variable dependiente.
  + La presencia de heterocedasticidad sugiere que, aunque las variables explicativas son relevantes, hay aspectos de la variabilidad en los datos que no están siendo capturados adecuadamente.

### Gráfico D: Residuals vs Leverage

* **Conclusión**:
  + El alto valor del estadístico F y el p-valor extremadamente bajo indican que el modelo en su conjunto es significativo. Sin embargo, el gráfico D puede ayudar a identificar observaciones con alto apalancamiento que podrían estar influyendo de manera desproporcionada en los resultados del modelo. Los puntos con alto apalancamiento y residuos estandarizados grandes indican observaciones influyentes que podrían estar afectando el ajuste del modelo.

Y que al 95% de confianza estas son las variables significativas: Destacan las presencias de los barrios, indicando que la presencia del alojamiento en alguno de ellos es influyente a la hora de determinar el precio de un anuncio.

## Análisis Clúster

Para determinar el número óptimo de clúster en que dividir el conjunto, se realiza el **Método del codo**, al no conocerse la cantidad de óptima de centroides *k* a utilizar de antemano (Delgado, 2018) => La enorme presencia de anuncios en Centro en comparación al resto de barrios provoca que el nº óptimo de clusters sea 1. Aun así, se realizan 4 en este estudio, y tras la división en clusters se han calculado las componentes principales:

Observando los gráficos de las varianzas, se aprecia que las dos primeras componenets principales explican un porcentaje aceptable de la variabilidad total de los datos. Más concretamente, **la componente principal 1 captura el 24,97% de la varianza total de los datos, y la componente principal 2 el 12,78%**. Por ello, se procede a continuar el análisis con ellas.

Las variables que más influyen en la primera componente principal son "*accommodates*" y "*beds*" (negativamente), así como "*price*" (también negativamente). Esto sugiere que la **CP1** puede estar representando un factor que combina elementos de capacidad y costo, donde una mayor capacidad de alojamiento y precio más alto se asocian con valores más bajos de **CP1.**

A su vez, las variables que más influyen en la segunda componente principal (**CP2**) son "*bathrooms*", "*room\_type*", y "*neighbourhood\_cleansed*" (todas negativamente). Esto sugiere que la **CP2** puede estar representando un factor relacionado con la calidad o el tipo de alojamiento y la ubicación, donde un mayor número de baños, ciertos tipos de habitaciones y ciertos vecindarios están asociados con valores más bajos de **CP2**.

Si se representan estas 2 primeras componentes principales, representando los clusters se tiene:

Ya que las variables más importantes de las variables con más peso en las componentes principales afectan todas negativamente, implica que aquellas con más mayor precio, pero a su vez mayor valores del resto (*accommodates, beds, bathrooms, room\_type. neighbourhood\_cleansed)*, es decir, los alojamientos más equilibrados, serán los que se encuentren en la zona más cercana al origen en los gráficos, por tanto, los alojamientos pertenecientes al primer clúster.

Aquellos alojamientos con mayor valor de la componente 2 y valor medio de la componente 1 se encuentran en los clústeres 2 y 3, es decir, serán aquellos con menor número de baños y peor tipo de habitación. También es posible que a su vez implique que el alojamiento se encuentre en **Centro**, al tener este barrio un valor bajo asociado y ser aquel con mayor oferta.

Finalmente, los alojamientos con altos valores en la componente 1 y valores bajos en la 2 serían aquellos recogidos en el clúster 4, es decir, aquellos con un número de baños considerable, tipo de habitación bueno y, probablemente no perteneciente a **Centro.** A su vez, serían alojamientos con pocos alojamientos y/o pocas camas. No serían alojamientos con precios bajos y el resto de características porque la tendencia que mostraban los datos por barrio en el análisis descriptivo inicial, sobre todo los outliers, es que los casos excepcionales se daban siempre por precios superiores o muy superiores. Se debe recordar como justificación a esto que no había ni un solo caso de outlier en el precio por barrio por precio bajo. Todos eran por exceso.

## Modelo Logit

La última parte del análisis es con la que se busca responder a la pregunta “¿Es el precio de un determinado anuncio adecuado o excesivo?”

Para ello, se ha desarrollado un modelo de elección binaria Logit tomando las variables que el modelo de regresión había devuelto como significativas. Los resultados son estos:

### Gráfico A: Deviance Residuals vs Fitted Values

* **Observaciones**:
  + La línea roja suavizada se curva alejándose de 0 hacia ambos extremos de los valores ajustados.
  + Hay una dispersión considerable en los puntos azules, especialmente en los extremos negativos de los valores ajustados.
  + Los residuales parecen estar dentro de las bandas de confianza (en gris) para la mayoría de los puntos, pero hay una dispersión más amplia en los valores extremos.
* **Análisis**:
  + La curvatura en la línea roja indica que puede haber una falta de linealidad en la relación modelada por el logit.
  + La mayor dispersión en los extremos sugiere que el modelo tiene dificultades para ajustar bien los valores extremos, lo que podría indicar la presencia de valores atípicos o una posible falta de alguna variable relevante en el modelo.

### Gráfico B: Standardized Residuals vs Leverage

* **Observaciones**:
  + La mayoría de los puntos están dentro de las bandas de confianza grises, pero hay algunos puntos fuera de estas bandas.
  + La línea roja suavizada se mantiene cerca de 0 en general, pero se eleva ligeramente hacia el extremo derecho del eje X.
  + Se observan varios puntos con altos residuales estandarizados y leverage, especialmente en los extremos.
* **Análisis**:
  + Los puntos con alto leverage y residuales estandarizados significativos podrían ser puntos de datos influyentes que afectan la estabilidad del modelo.
  + La elevación en la línea roja hacia el extremo derecho sugiere que estos puntos influyentes están teniendo un impacto en la estimación del modelo, lo que puede indicar que algunos datos están teniendo una influencia desproporcionada.
  + Es necesario investigar estos puntos influyentes para determinar si son errores, valores atípicos o si hay alguna característica importante del modelo que no se está capturando adecuadamente.

### Gráfico C: Receiver Operating Characteristic (ROC) Curve

* **Observaciones**:
  + La curva ROC está claramente por encima de la línea diagonal, indicando que el modelo tiene un poder de discriminación razonable.
  + El área bajo la curva (AUC) es 0.77.
* **Análisis**:
  + Un AUC de 0.77 indica que el modelo tiene una buena capacidad para distinguir entre las clases positivas y negativas, aunque hay margen para mejorar.
  + La curva ROC muestra que el modelo logra una alta tasa de verdaderos positivos con una tasa de falsos positivos razonablemente baja.

### Resumen del análisis

* **Gráfico A**: La curvatura en la línea roja y la dispersión en los extremos sugieren problemas potenciales de ajuste en los extremos de los valores predichos.
* **Gráfico B**: La presencia de puntos con alto leverage y residuales significativos indica posibles puntos influyentes que deben ser investigados más a fondo.
* **Gráfico C**: El AUC de 0.77 muestra que el modelo tiene una buena capacidad de clasificación, pero no es perfecto y podría beneficiarse de ajustes adicionales.

## Conclusiones

### Relativas al análisis descriptivo

* El mayor número de anuncios se da en **Centro**, barrio que acumula la mayor parte de la oferta.
* El rango más común de precios por noche es **]0, 250]**.
* De las variables que determinan el precio de los anuncios, ***bathroom*** y ***room\_type*** son las dos con más peso.
* Por tanto, los tres principales factores que determinan el precio por noche de un alojamiento turístico no hostelero son: **El barrio en que se encuentra, el número de baños y el tipo de habitación que se oferta.**

### Relativas al análisis Clúster

* El efecto de **Centro** eclipsa el efecto del resto de los barrios, haciendo que el número óptimo de clusters sea 1 pese a que se calculen 4.
* La primera componente principal, **CP1**, relacionada con la capacidad y el costo del alojamiento, muestra que la variedad en el tipo de alojamiento, así como su capacidad de alojamiento y camas disponibles son factores claves en las preferencias de los usuarios.
* La segunda componente principal, **CP2**, relacionada con la calidad de los alojamientos y su ubicación, muestra que el número de baños, los tipos de habitaciones de los alojamientos y el barrio en que se encuentra el alojamiento influyen a su vez también las preferencias de los usuarios.

### Relativas al modelo Logit

El modelo generado es bueno para identificar alojamientos con precios excesivos.

El análisis de las variables muestra que:

* Las variables como el número de plazas disponibles ("*accommodates*"), el número de baños ("*bathrooms*") y el tipo de propiedad ("*property\_type*") están asociadas con un aumento en la probabilidad de precios altos, siendo el número de baños la más importante, implicando que el número de baños de un alojamiento es el factor más importante en el mercado de alquiler de alojamientos turísticos no hosteleros de Málaga de la plataforma Airbnb.
* Las variables como la verificación de identidad del anfitrión ("*host\_identity\_verified*"), el tipo de habitación ("*room\_type*") y el número mínimo de noches ("*minimum\_nights*") tienen coeficientes negativos, por lo que tienden a estar asociadas con una menor probabilidad de que el precio del alojamiento sea alto. En particular, el tipo de habitación tiene el mayor impacto negativo, por lo que se infiere que el tipo de habitación que se oferta en los anuncios es también muy determinante para determinar si el precio es excesivo o no.

#### Limitaciones del modelo

* Sensibilidad baja El modelo tiene dificultades para identificar correctamente todos los casos positivos.
* Alta especificada El modelo es muy bueno para identificar los casos negativos.

### Posibles usos del modelo

* **Optimización de Estrategias de Precios:** Las plataformas de alquiler de alojamientos pueden utilizar este modelo para identificar propiedades con potencial para precios más altos. Esto puede ayudar a los anfitriones a ajustar sus precios de manera más competitiva y maximizar sus ingresos, basándose en las características significativas identificadas por el modelo, y por el contrario, desde el punto de vista del cliente, puede utilizar este modelo para comprobar si el precio por noche de un alojamiento de su interés es excesivo o no, ayudando en la toma de decisión de alquilar el alojamiento.
* **Segmentación de Mercado:** Las agencias de marketing y las plataformas de alquiler pueden segmentar mejor sus mercados utilizando las predicciones del modelo. Por ejemplo, pueden dirigir campañas de marketing específicas a propiedades con características que se correlacionan positivamente con precios altos, como un mayor número de baños y tipos específicos de propiedades.
* **Evaluación de Nuevas Propiedades:** Los nuevos anfitriones pueden usar el modelo para evaluar el potencial de sus propiedades en términos de precio. Al entender qué características aumentan la probabilidad de precios altos, los anfitriones pueden invertir en mejoras específicas, como aumentar el número de baños o cambiar el tipo de propiedad, para aumentar el valor percibido de su alojamiento.
* **Mejoras en la Plataforma:** Las plataformas pueden utilizar los resultados del modelo para mejorar sus algoritmos de recomendación y búsqueda, destacando propiedades que tienen una alta probabilidad de ser listadas a precios más altos. Esto puede mejorar la experiencia del usuario al proporcionar recomendaciones más precisas y relevantes.