

Tu mejor airbnb en 10 ciudades europeas.

* Castillo Reynoso Ivan Alexis, * Loreto Silva Marco,
* Rodriguez Martínes Fanny Arlin, * Rubio Juárez Elizabeth

Temas Selectos de Física Computacional II.

Universidad Nacional Autónoma de México, Facultad de Ciencias, México

5 de junio de 2023

Resumen.

En esta investigación se analizó el dataset **Airbnb Prices in European Cities** [1] con el fin de brindar un panorama general de las opciones de Airbnb que se ofertan en Ámsterdam, Atenas, Barcelona, Berlin, Budapest, Lisboa, Londres, Paris, Roma y Vienna. Se estudiaron los factores más relevantes en el precio del Airbnb y la satisfacción del huésped. Se observó una variación en el precio promedio de los alojamientos de Airbnb entre los días de semana y los fines de semana. También se encontró que la limpieza está altamente relacionada con la calificación que los huéspedes asignan al Airbnb. Así mismo, se presentan mapas con las ubicaciones de los Airbnb en cada ciudad, clasificados por precio y cercanía a las principales atracciones turísticas.

Introducción

Al momento de planear un viaje, siempre buscamos la mejor opción. Para ello se consideran varios aspectos, pero en particular el precio del lugar en el cual nos vamos a alojar. Actualmente, airbnb es la plataforma más usada para rentar casas, departamentos, cabañas, entre otros. De acuerdo a Airbnb, una media de dos millones de viajeros se alojan cada noche en lugares proporcionados a través de la plataforma. [2]

Es por ello que en este proyecto te presentamos un análisis de los airbnb's de 10 ciudades europeas. Lo anterior mediante el dataset en Kaggle llamado **Airbnb Prices in European Cities**.

Dataset

Este dataset se obtuvo mediante un experimento automatizado basado en web-scraping. Con el uso de un marco de automatización web (Selenium WebDriver), se ejecutaron consultas de búsqueda en la plataforma de Airbnb que se referían a alojamientos en 10 ciudades europeas importantes para dos personas y dos noches. Las ofertas se recopilaron de cuatro a seis semanas antes de las fechas de viaje, y los precios recopilados se refieren al monto total adeudado por el alojamiento, incluida la tarifa de reserva y la tarifa de limpieza. Para cada ciudad se prepararon dos conjuntos de datos, incluyendo ofertas para los días de la semana (martes-jueves) y fines de semana (viernes-domingo). [2]

Ambos conjuntos de datos poseen los siguientes atributos:

- **Unnamed:0 (int64).** Nos da el ID del airbnb.
- **realSum (float64).** Monto total adeudado por el alojamiento, incluida la tarifa de reserva y la tarifa de limpieza.
- **room_type (object).** Tipo de habitación (shared room, private room, entire home/apt).
- **room_shared (bool).** Dummy para habitaciones compartidas.
- **room_private (bool).** Dummy para habitaciones privadas.

- **person_capacity (float64)**. Número máximo de visitantes.
- **host_is_superhost (bool)**. Se refiere a si el anfitrón es superhost o no.
- **multi (int64)**. Dummy para lugares ofrecidos por anfitriones con 2–4 lugares.
- **biz (int64)**. Dummy para lugares ofrecidos por anfitriones con más de 4 lugares.
- **cleanliness_rating (float64)**. Reseña de los visitantes con escala que llega a 10.
- **guest_satisfaction_overall (float64)**. Reseña de los visitantes con escala que llega a 100.
- **bedrooms (int64)**. Número de habitaciones.
- **dist (float64)**. Distancia al centro de la ciudad en Km.
- **metro_dist (float64)**. Distancia a la estación de metro más cercana en Km.
- **attr_index (float64)**. Índice de atracción (cosas que hacer).
- **attr_index_norm (float64)**. Índice de atracción normalizado. Su escala va hasta 100.
- **rest_index (float64)**. Índice de restaurantes.
- **rest_index_norm (float64)**. Índice de restaurantes. Su escala va hasta 100.
- **lng (float64)**. Longitud del lugar.
- **lat (float64)**. Latitud del lugar.

El attr_index para el Airbnb con índice j basado en K puntos de interés o atracciones turísticas es calculado como

$$attr_index_j = \sum_{k=0}^K \frac{R_k}{d_{jk}}$$

donde R es el número de reseñas para la atracción k y d_{jk} es la distancia entre el Airbnb j y la atracción turística k . De manera similar se define el rest_index.

Regresión lineal

Es una técnica de aprendizaje automático supervisado que se utiliza para predecir una variable de salida continua (y) a partir de una o varias variables de entrada (x).

Existen dos tipos de regresión lineal:

1. **Simple**. Se utiliza una única variable de entrada (x) para predecir la variable de salida (y).
2. **Múltiple**. Se utilizan varias variables de entrada (x_1, x_2, \dots, x_n) para predecir la variable de salida (y).

En particular, utilizaremos la regresión lineal múltiple.

Se expresa como $y = b_0 + b_1x + \epsilon$ donde y es la variable de salida, x es la variable de entrada, b_0 y b_1 son los coeficientes de regresión y ϵ es el término de error. Se busca encontrar los valores de b_0 y b_1 que minimicen la suma de los errores al cuadrado.

Dentro de los métodos para encontrar los valores de b_0 y b_1 que minimicen la suma de los errores al cuadrado está el método de mínimos cuadrados. Este método consiste en calcular la suma de los errores al cuadrado para cada valor de b_0 y b_1 y elegir los valores que dan como resultado la menor suma de errores al cuadrado.

Otro método es el descenso del gradiente (el cual utilizaremos). Es un algoritmo iterativo de primer orden que actualiza todos los parámetros a modo de que el error se minimice de forma local.

Los supuestos de la regresión lineal son:

1. **Relación lineal**
2. **Independencia**

3. Homocedasticidad
4. Normalidad

Coeficiente de determinación R^2

Es una medida de la calidad del modelo de regresión, nos indica el porcentaje de los datos que se ajustan a la regresión lineal. Se calcula como la proporción de la varianza de y .

Escalado de características

El descenso del gradiente es una técnica buena para aproximarse al valor mínimo del Error. Sin embargo, si los valores de las variables dependientes son de diferentes magnitudes, éste podría fallar.

Para evitar dicho fallo debemos normalizar los rangos a modo de que la contribución de las características sea proporcional. Existen al menos tres métodos de normalización: re-escalado, estandarización y magnitud unitaria. En particular, utilizaremos la estandarización:

$$x' = \frac{x - \bar{x}}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x^{(i)} - \bar{x})^2}}$$

Regularización

Es una técnica utilizada en el aprendizaje automático para prevenir el sobreajuste de un modelo.

Esta agrega una penalización a los parámetros del modelo que pueden tomar valores extremos, lo que reduce la complejidad del modelo y lo hace menos propenso a sobreajustarse.

Hay dos tipos principales de regularización:

1. **Regularización L1.** También es conocida como LASSO, agrega una penalización a la suma de los valores absolutos de los coeficientes del modelo. Es decir, los coeficientes del modelo pueden tomar valores cercanos a cero, lo que resulta en la eliminación de características no importantes.

$$+ \lambda \sum_{j=0}^p |w_j|$$

La cantidad de penalización se controla mediante un hiperparámetro denominado parámetro de regularización. Cuanto mayor sea el valor del parámetro de regularización, mayor será la penalización y menor será la complejidad del modelo.

2. **Regularización L2.** También es conocida como regresión de Ridge, agrega una penalización a la suma de los valores cuadrados de los coeficientes del modelo. Esto significa que los coeficientes del modelo no pueden tomar valores extremos y deben estar cerca de cero, lo que reduce la complejidad del modelo.

$$+ \lambda \sum_{j=0}^p |w_j|^2$$

La cantidad de penalización se controla mediante un hiperparámetro denominado parámetro de regularización. Cuanto mayor sea el valor del parámetro de regularización, mayor será la penalización y menor será la complejidad del modelo.

Objetivo General

Se analizará el dataset **Airbnb Prices in European Cities** [1] con el fin de brindar un panorama general de las opciones de Airbnb que se ofertan en Ámsterdam, Atenas, Barcelona, Berlin, Budapest, Lisboa, Londres, Paris, Roma y Vienna. Se pretende determinar los factores más relevantes en el precio del Airbnb y la satisfacción del huésped, para así ubicar la mejor opción de alojamiento para el cliente durante su estadía.

Hipótesis

- El precio del Airbnb varía dependiendo si el alojamiento es entre semana o en fin de semana.
- El precio del Airbnb aumenta si se encuentra cerca de alguna atracción turística.
- Airbnb es la opción más flexible para viajes en pareja, amigos y grupos grandes.
- La limpieza es un factor determinante en la calificación que el huésped asigna al Airbnb.

Análisis y discusión.

Primer paso en la búsqueda del mejor Airbnb

Con el propósito de presentar un panorama general de la oferta de Airbnb, se realizaron mapas con las ubicaciones de los Airbnb para cada ciudad (Figura 5). Los precios se dividieron en 4 categorías delimitadas por los cuartiles y se muestran en los mapas en diferentes colores. Este proceso se repitió para el *attraction index*, las diferentes categorías se presentan en diferentes figuras.

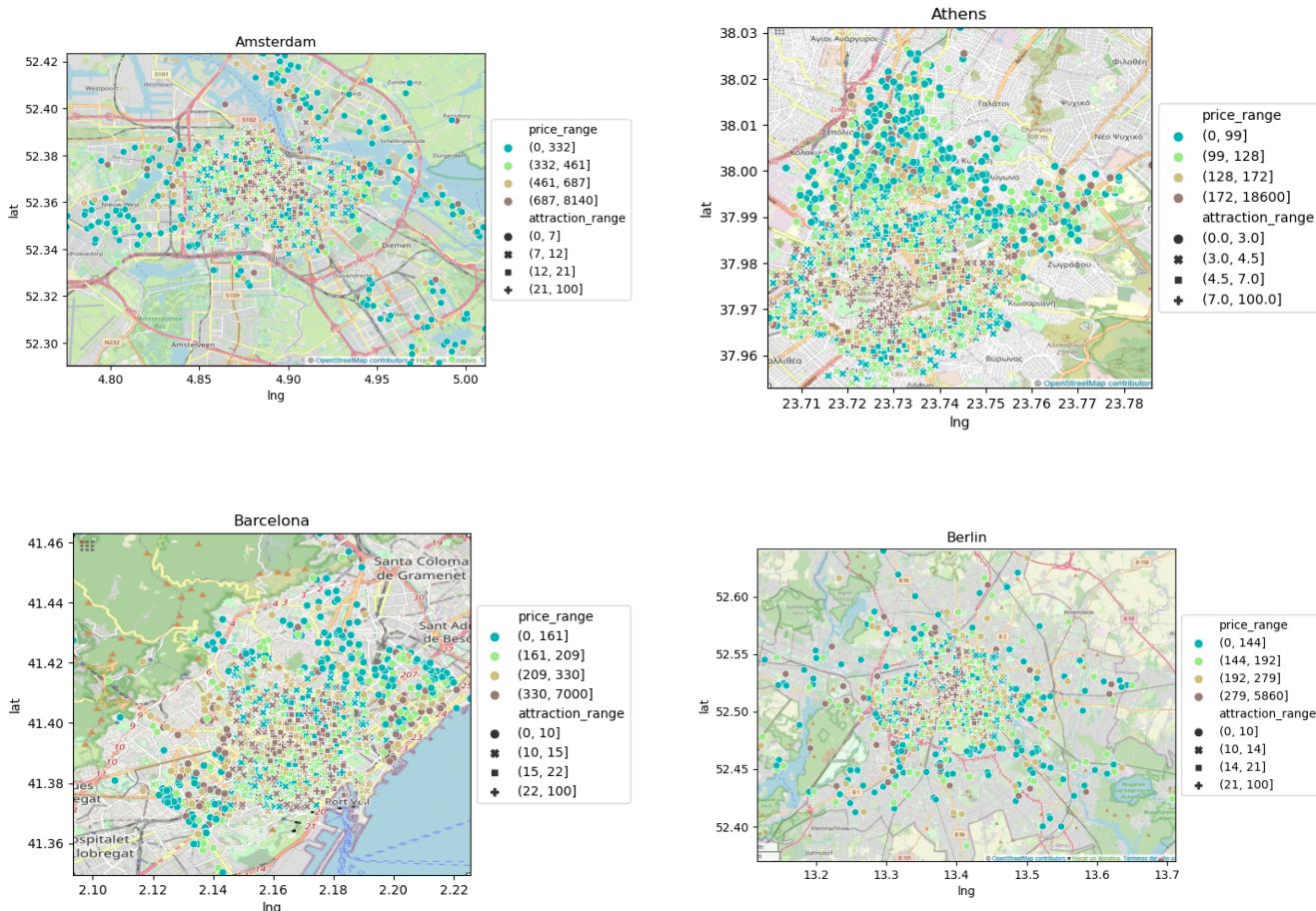




Figura 5: Mapas de las ubicaciones de Airbnb en cada ciudad

En la mayoría de las ciudades, los precios más altos se concentran alrededor de las principales atracciones, por ejemplo en Londres o en Roma. Sin embargo, es posible identificar algunos Airbnb que cuentan con un *attraction index* alto y un precio en la categoría más baja, como en París, Lisboa, Budapest, Barcelona, Berlín o Ámsterdam.

Se observa que en ciudades como Ámsterdam, Barcelona o Lisboa, las principales atracciones se encuentran distribuidas por toda la ciudad; por lo que es posible encontrar un Airbnb que esté cerca de alguna atracción y elegir entre una variedad de precios. Por otro lado, en ciudades como Roma, Londres y Atenas, las atracciones y los precios más altos se concentran alrededor del centro de la ciudad. En estas ciudades los precios decrecen a las afueras de la ciudad.

En París y Londres existe una cantidad significativamente mayor de Airbnb respecto a las otras ciudades del estudio, mientras que en Budapest el número de Airbnb es notablemente menor.

Características de los Airbnb

Complementando en cuanto a las características que los Airbnb de cada ciudad ofrecen, en la figura 6, podemos observar la distribución en la capacidad de los Airbnb de cada ciudad. Aspecto que nos interesa, si es que se busca viajar con varias personas.

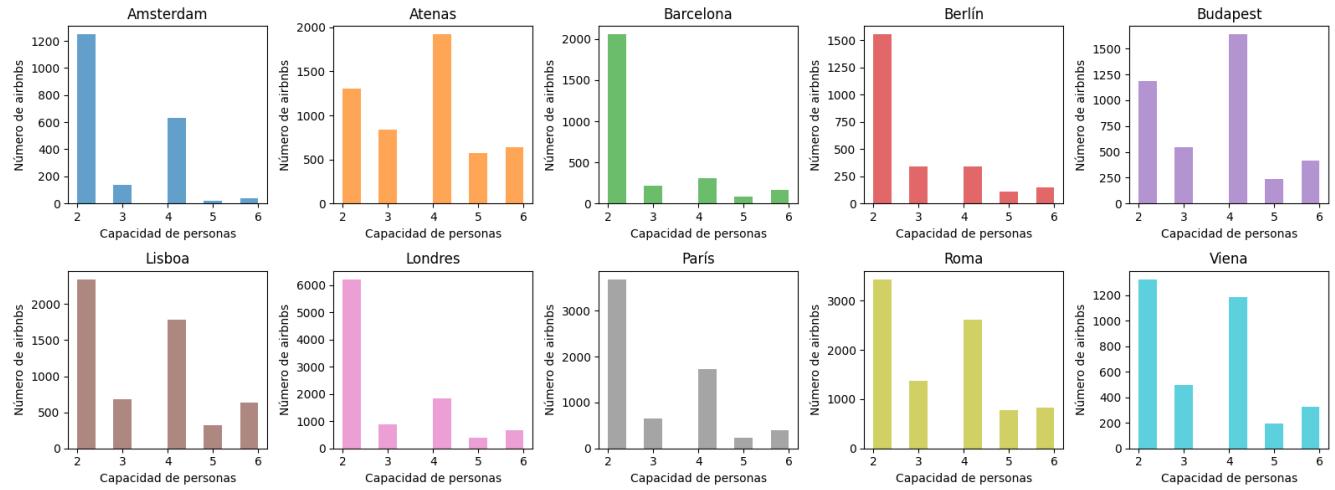


Figura 6: Histogramas de la capacidad de los Airbnb por ciudad

De estas gráficas, observamos que la mayoría de cuartos en Airbnb están adecuados para dos personas, siendo las opciones para viajar con más personas, Atenas, Lisboa, Budapest, Roma y Viena, mientras que Berlín y Barcelona tienen un enfoque a estancias de dos personas.

Lo cual es un factor a considerar si se planea hacer el hospedaje en Airbnb, ya que podemos ver un cierto mercado dependiendo de las ciudades.

Otra característica de interés son los tipos de Airbnb que se ofertan, en la figura 7 se presentan por ciudad.

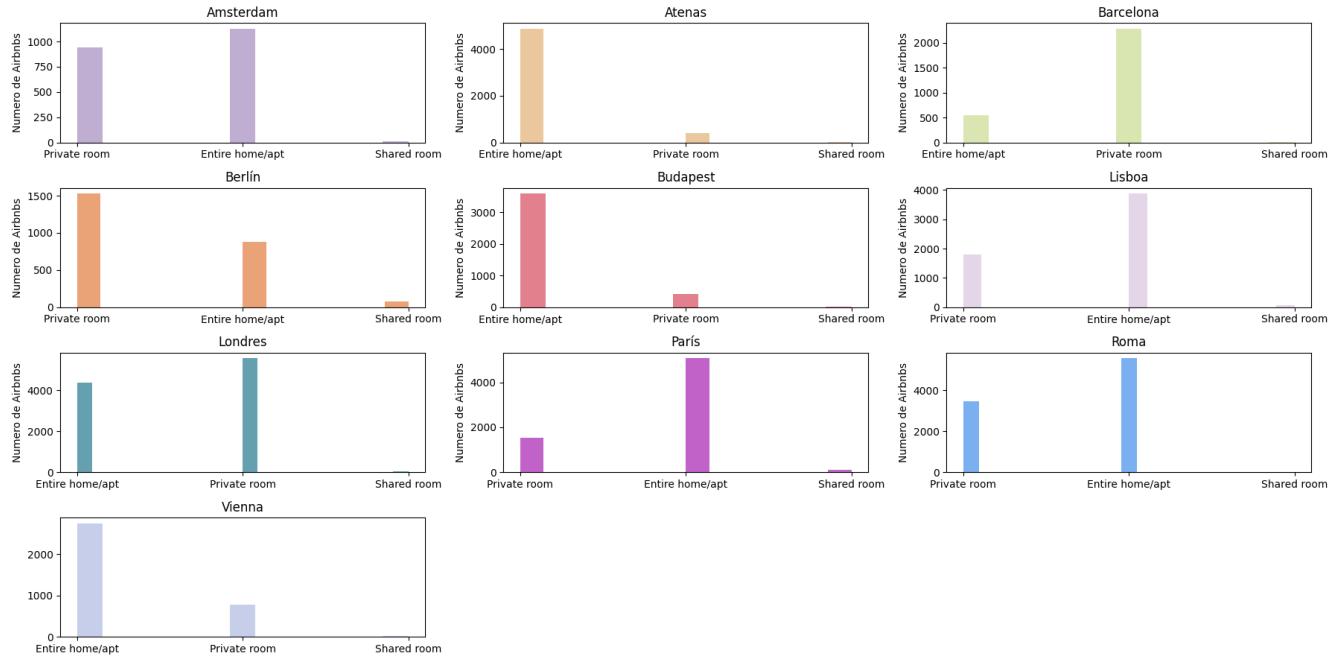


Figura 7: Histogramas de los tipos de Airbnb ofertados en cada ciudad.

De los histogramas, se puede observar que en general los cuartos privados y las casas enteras/apartamentos son el tipo de Airbnb más popular que se ofrecen, en general, los cuartos compartidos no son ofertados, esto indica que no son populares. Por otra parte, en Amsterdam y Londres se observa que la cantidad de cuartos privados y casas enteras/apartamentos son muy similares, sin embargo, en Amsterdam son más populares las casas enteras/apartamentos y en Londres los cuartos privados. En las demás ciudades se observa que hay una diferencia mayor entre los tipos de Airbnb, en Atenas, Budapest, Lisboa, París, Roma y Vienna son más populares las casas enteras/apartamentos, y en Barcelona y Berlin son más populares los cuartos privados. Se observa una tendencia a que en Barcelona y Berlin el enfoque de los Airbnb son para un número de personas reducido.

Críticas

Para conocer a las ciudades que cuentan con las mejores críticas o reseñas, se realizó un boxplot con la distribución de la variable *guest satisfaction overall*, esto se muestra en la figura 8.

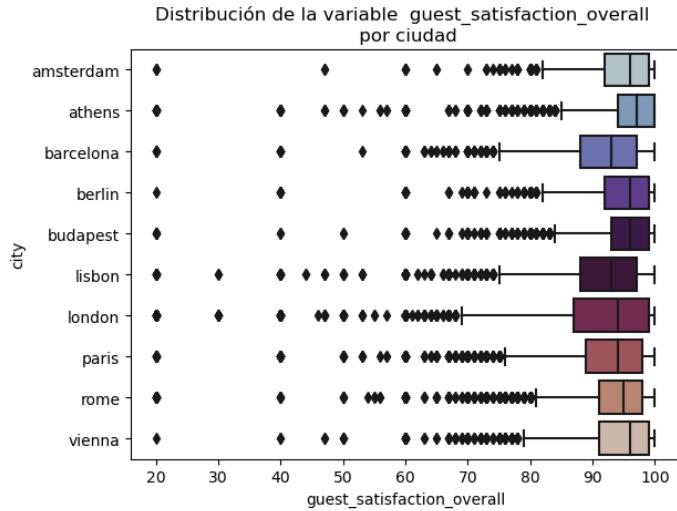


Figura 8: Distribución de críticas por ciudad

Se observa que, en general, la mediana de las distribuciones se concentra en puntuaciones entre 90 y 100 con la presencia de algunos outliers en valores menores a 70. Todas las distribuciones son asimétricas a la izquierda, por lo que la mayoría de críticas son buenas o excelentes. Para Barcelona y Lisboa, el tercer cuartil se encuentra cercano a 95, indicando que el 75 % de los Airbnb tienen a lo más una puntuación dé 95. Para el resto de los destinos, el tercer cuartil es más alto, siendo Atenas el mayor.

Se observa que para Londres el rango intercuartílico es mayor que para cualquier otro destino, lo que representa que existe un mayor número de puntuaciones más bajas que en otras ciudades.

Continuando con las reseñas, se observará la relación entre las críticas y otras de las variables, en la figura 9

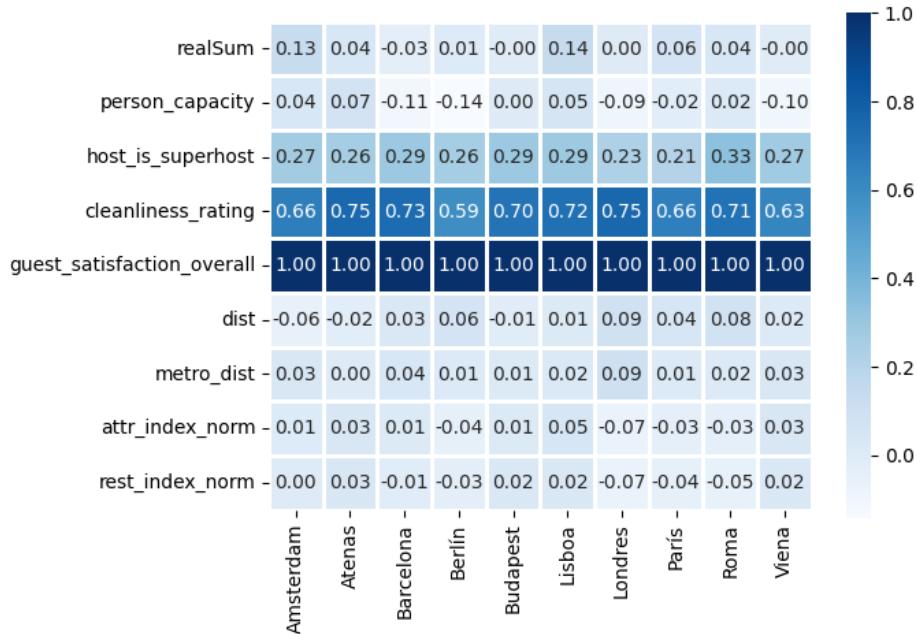


Figura 9: Distribución de críticas por ciudad

Se observa que los que están relacionados con las críticas son la limpieza y en menor medida si un host es superhost. En el caso de la calificación que se le da a la limpieza esta relacionada aproximadamente entre 0.59 y 0.75; por otra parte, la otra característica relacionada es si un host es superhost, sin embargo, esta relación está entre 0.21 y 0.33, es decir, es baja la relación, a pesar de esto se analizarán estas dos características.

Ahora, se observará la relación entre las críticas y la limpieza, para esto se presenta la figura 10.

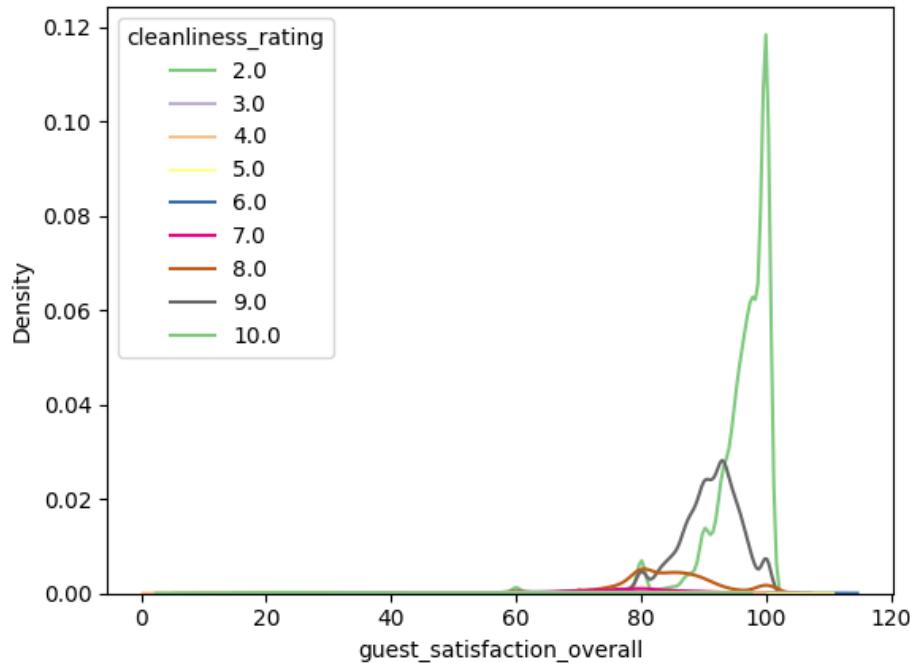


Figura 10: Distribución de críticas por ciudad

Se observa que en general la limpieza es buena, también se tiene que hay una relación entre las críticas y la limpieza, si se tiene buena limpieza las críticas son mejores, y mientras más sucio este las críticas son menos buenas.

La otra variable a analizar es si el host es un superhost y su relación con las reseñas, para esto se presenta la figura 11.

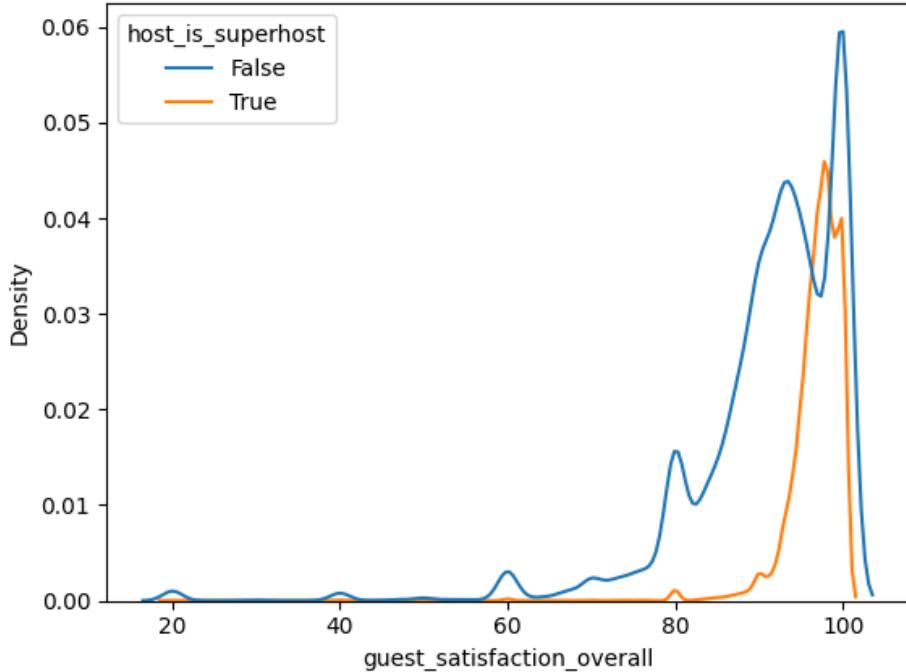


Figura 11: Distribución de críticas por ciudad

Se observa que para el caso de que el host si es un superhost las criticas tienden al mayor puntaje de críticas, en el caso de que no lo son se observa que el puntaje esta más disperso, por tanto podemos decir que si se puede relacionar en una pequeña medida que si el host es superhost con las críticas buenas.

Precios

Con el objetivo de identificar los destinos más atractivos para cada presupuesto y necesidades. Se buscó comparar los precios promedios de los Airbnb en cada ciudad para estancias entre semana y durante los fines de semana.

Obteniendo así el gráfico de la figura 12.

De este gráfico podemos observar una clara diferencia en los precios en Amsterdam y el resto de ciudades, siendo la segunda más cara, París y Londres como la tercera más cara.

A diferencia de la más accesible en cuanto a precios, tenemos a Atenas, con casi 400 euros de diferencia.

De la misma gráfica encontramos que en 7 de las 10 ciudades, los precios se incrementan durante los fines de semana, siendo de nuevo Amsterdam en caso donde esta diferencia es más grande. Siendo esta información a considerar en cuanto a la planeación de un viaje a una de estas ciudades con alojamiento en un airbnb.

Por otra parte, dado que realizaremos una regresión lineal teniendo como variable de salida *realSum*; es de relevancia conocer la relación que tiene con las otras variables con un heatmap.

De acuerdo al heatmap, la variable *realSum* (precio) tiene mayor relación con las variables *person_capacity*

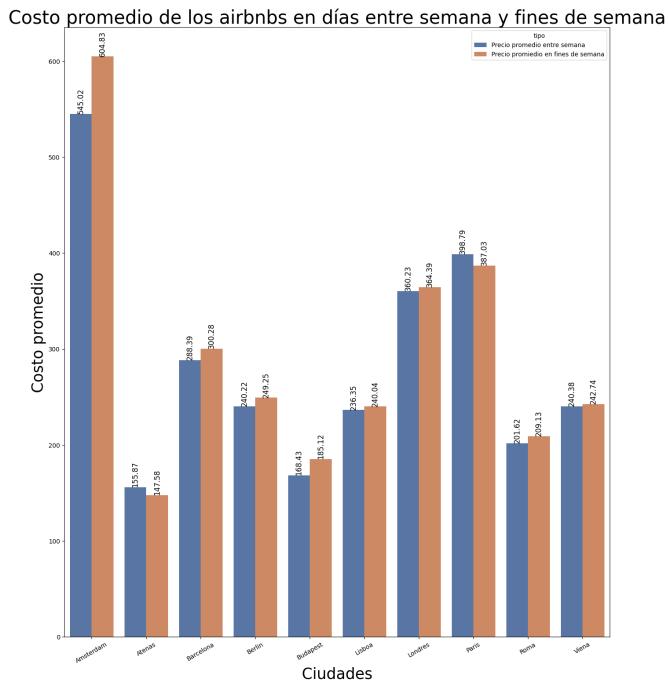


Figura 12: Costo promedio en fines de semana y días entre semana

y *bedrooms* en todas las ciudades. Es de relevancia mencionar que solo se consideraron las primeras cinco variables que se relacionaban más en cada ciudad.

Con estas variables se realizó la regresión lineal utilizando la variable *people_per_bedrooms*. Se obtuvieron los coeficientes de determinación, los cuales indican que solo 23.78 % de los datos de entrenamiento se ajustan a la regresión lineal y para los datos de prueba solo el 24.11 %.

La ecuación obtenida es la siguiente:

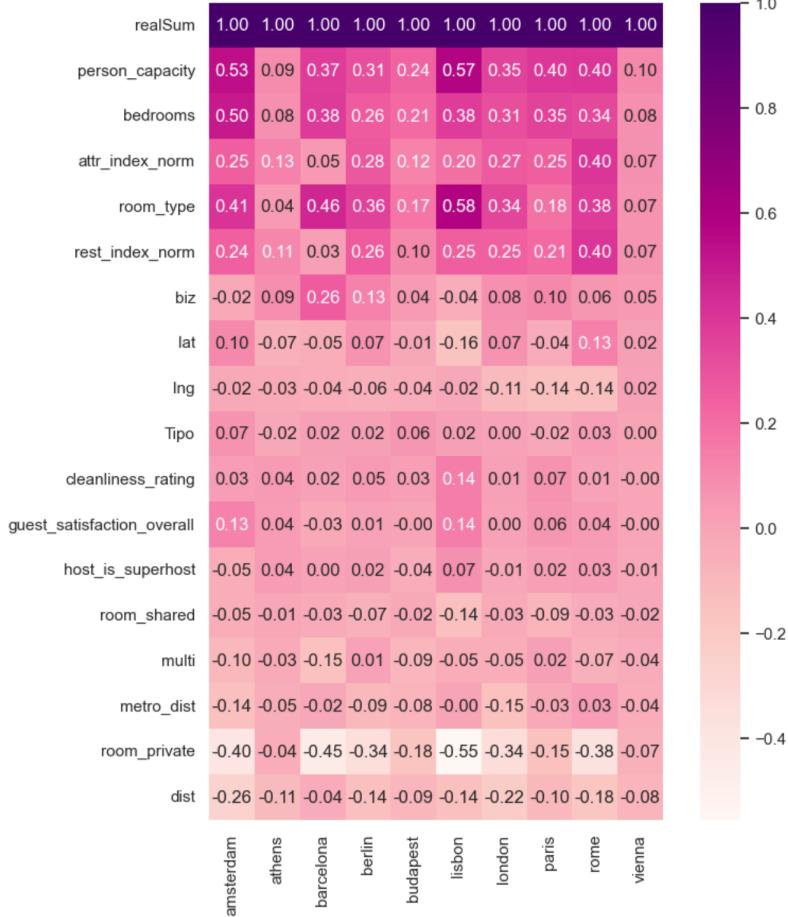


Figura 13: Relación de la variable *realSum* con las demás.

$$y = -0.1505 + 0.1954x_1 + 0.1579x_2 + 0.1055x_3 + 0.01242x_4$$

donde $y = \text{realSum}$, $x_1 = \text{attr_index_norm}$, $x_2 = \text{bedrooms}$, $x_3 = \text{person_capacity}$ y $x_4 = \text{rest_index_norm}$.

Además, se predijo para calcular el error (MSE) encontrando que estos son demasiado altos, por lo que, entrenar el modelo con esas variables no fue óptimo. Lo anterior nos hace reflexionar en que sería una mejor opción hacer la regresión lineal para cada ciudad en lugar de hacerlo para todas. Dado que, el porcentaje de relación varía en cada caso.

Por otra parte, se realizó la regresión lineal a través del descenso del gradiente (SGD) obteniendo que 23.74 % de los datos se predicen. De manera gráfica se tiene lo siguiente:

De la gráfica anterior se concluye que estos conjuntos no se encuentran tan relacionados debido a que la taza de aprendizaje no es alta.

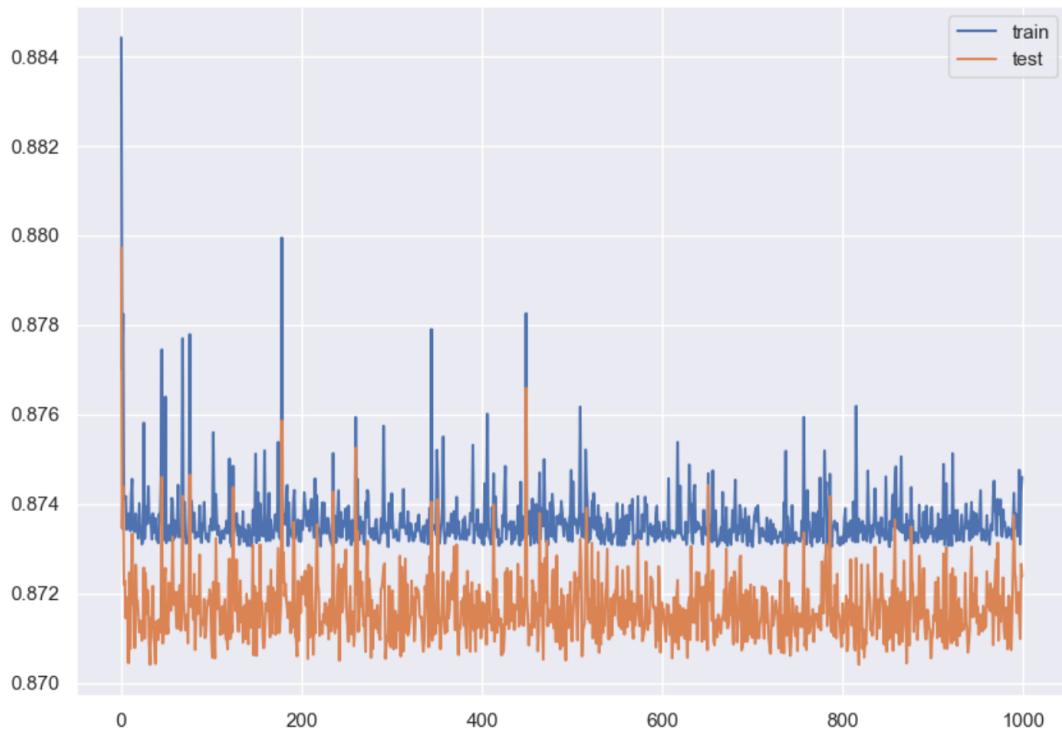


Figura 14: Conjunto de entrenamiento (train) y prueba (test).

Consecuentemente, se realizaron gráficas de tres de los cuatro supuestos de la regresión lineal para encontrar cual no se cumple.

Normalidad

Del histograma 15 se observa que claramente no se tiene normalidad en los datos y es una de las razones por las cuales al implementar la regresión lineal no se obtuvieron resultados óptimos.

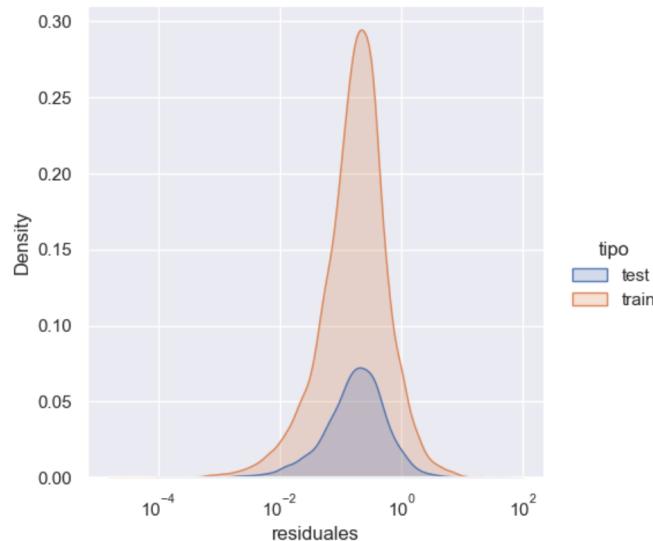


Figura 15: Gráfica de normalidad.

Homocedasticidad

De la gráfica 16 tenemos que el error cambio en todos los valores de la variable independiente (*people_per_bedrooms*).

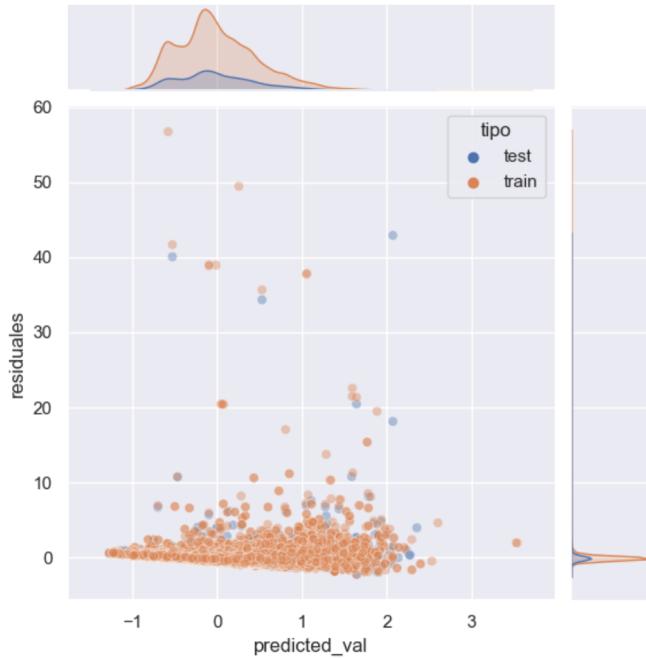


Figura 16: Gráfica de homocedasticidad.

No autocorrelación

De la gráfica 17 tenemos que no se tiene autocorrelación entre los valores del término de error.

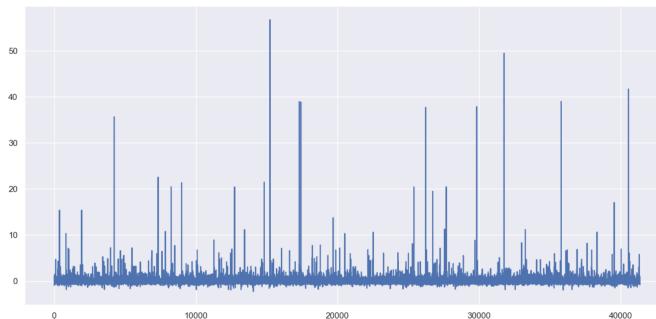


Figura 17: Gráfica de no autocorrelación.

Recomendaciones

Con toda la información obtenida con anterioridad, y de acuerdo a las preferencias de cada cliente. Presentamos un ejemplo de como encontrar la mejor opción en Lisboa para un viaje de pareja y de tres o más personas.

Priorizando el precio bajo, la cercanía con la ciudad, el metro y los atractivos como los restaurantes y atracciones en general. Obtuvimos los alojamientos mostrados en la tabla 1.

Tabla 1: Recomendaciones de hospedaje en Lisboa

ID	Precio	Capacidad de personas	Distancia a la ciudad	Distancia al metro	Índice de atracciones	Índice de restaurantes	Tipo	Preferencia
2934	262.664	2	1.029	0.41258	3031.840	827.819	FN	Pareja
90	263.602	2	1.0293	0.4126	3028.99	828.072	ES	Pareja
3756	342.167	2	0.41651	0.39818	388.545	1489.62	FN	Pareja
3776	323.874	2	0.17084	0.20971	490.709	1431.82	FN	Pareja
1121	298.546	2	0.40927	0.35126	367.998	1480.01	ES	Pareja
95	374.296	4	1.04332	0.41167	1527.12	723.681	ES	Familia
894	263.602	4	0.29587	0.30614	512.286	1620.22	ES	Familia
3633	465.056	6	0.36436	0.35279	405.036	1617.13	FN	Familia
790	188.086	4	0.3969	0.39256	387.799	1625.48	ES	Familia
3688	355.066	5	0.46869	0.45758	375.242	1599.19	FN	Familia

Con esta información podemos ubicar en el mapa de Lisboa los alojamientos, tal y como se muestra en la figura 18.

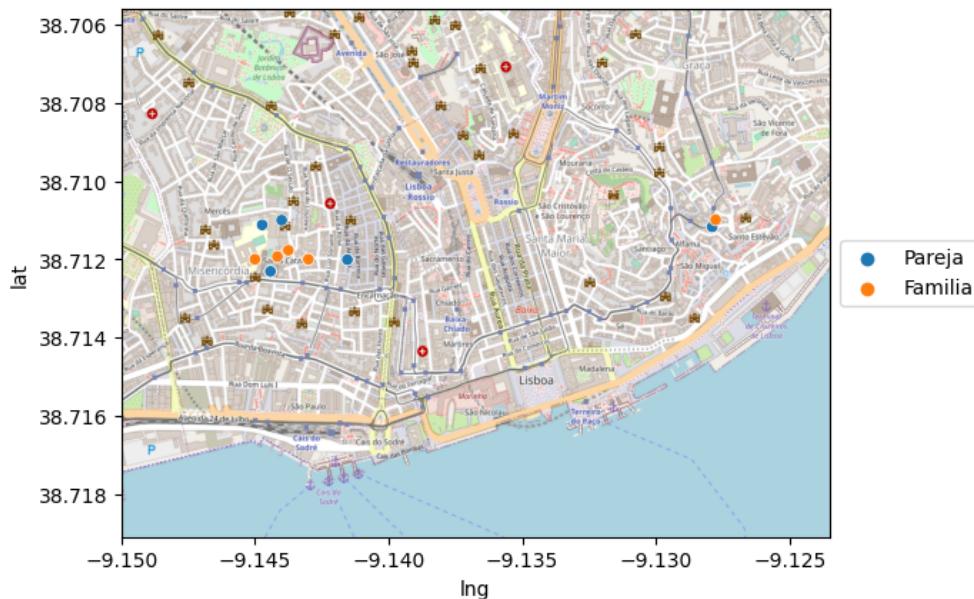


Figura 18: Ubicación de las recomendaciones para alojamientos en Lisboa

Con esta gráfica, podemos fácilmente identificar la ubicación de cada estancia, siendo los puntos verdes, las estancias para parejas y los azules para estancias de 3 o más personas.

Con lo que sumado a la información encontrada en las secciones previas, podemos encontrar la mejor opción en cuanto a ciudad y posteriormente encontrar las estancias que ofrecen lo que buscamos, ya sea priorizando el precio, las atracciones, entre otros aspectos.

Conclusiones

Los precios más altos de los alojamientos de Airbnb en ciudades como Londres y Roma se concentran alrededor de las atracciones principales en el centro de la ciudad. En cambio, en ciudades como Lisboa y Barcelona, las atracciones están distribuidas por toda la ciudad, lo que permite encontrar alojamientos cerca de diferentes atracciones a una variedad de precios.

La mayoría de los alojamientos de Airbnb están diseñados para estancias de dos personas, lo cual limita la oferta para grupos mayores a tres. Sin embargo, se identificaron ciudades con una mayor oferta para viajes en grupos grandes.

En la mayoría de las ciudades, los alojamientos más comunes en Airbnb son casas completas o apartamentos, mientras que los cuartos compartidos no son comunes.

Las calificaciones de los alojamientos de Airbnb están relacionadas principalmente con la limpieza. Además, se encontró que los hosts con la designación de "superhost" tienden a tener calificaciones más altas. Sin embargo, la relación entre estas variables indica que si un host es un superhost, es más probable que tenga una buena calificación, pero no garantiza una calificación alta.

Se observó una variación en el precio promedio de los alojamientos de Airbnb entre los días de semana y los fines de semana. Amsterdam resultó ser la ciudad más cara en ambos casos, mientras que Atenas fue la más económica.

La variable "*realSum*" muestra una mayor relación con las variables "*personCapacity*" y "*bedrooms*".^{en} en todas las ciudades analizadas. Esto sugiere que la capacidad de personas y el número de habitaciones influyen en el precio real de los alojamientos.

A través del análisis de regresión lineal, no se encontró una relación lineal significativa entre las variables de entrada *personCapacity* y *bedrooms* y la variable de salida *realSum*. Los coeficientes de determinación indican que solo un pequeño porcentaje de los datos se ajusta a la regresión lineal, tanto en los datos de entrenamiento como en los de prueba (23.78 % y 24.11 %, respectivamente). Esto sugiere que otros factores pueden tener una mayor influencia en el precio real de los alojamientos. Además, dos de los cuatro supuestos de la regresión lineal no se cumplen.

Referencias

¹K. Gyódi and L. Nawaro, Airbnb prices in european cities, **2023**.

²K. Gyódi and L. Nawaro, «Determinants of airbnb prices in european cities: a spatial econometrics approach», *Tourism Management* **86**, 104319 (2021).