Reporte Ejecutivo

Actividad 7

Evidencia 2. Análisis y reporte ejecutivo de indicadores

Alumnas:

Miroslava Arredondo

Sarahí Gómez

Alexa Gómez

Mariana Torres

Docentes:

Mtra. María Luisa Gómez Barrios

Dr. Alfredo García Suárez

La forma en que las personas encuentran alojamiento temporal en todo el mundo ha sido cambiada por Airbnb, que ha creado un mercado que combina elementos de la economía colaborativa con la industria hotelera tradicional. Este análisis analiza los datos proporcionados por Airbnb de dos ciudades europeas: Barcelona, un epicentro turístico en España, y Atenas, una ciudad llena de historia y cultura en Grecia. Aunque Airbnb atrae millones de turistas a ambas ciudades cada año, sus mercados de alquiler presentan diferencias que son esenciales para comprender la dinámica de la oferta y la demanda en cada uno.

Propuesta de Solución

Varios pasos clave se utilizaron para desarrollar el análisis de datos de Airbnb para Barcelona y Atenas:

Para garantizar que el análisis fuera sólido y confiable, la depuración de valores nulos se centró en eliminar o imputar datos faltantes. Se utilizó un método metódico: cada columna con muchos valores nulos fue revisada individualmente para determinar si podían ser imputadas o si debían eliminarse del análisis. Dependiendo del tipo de datos, se utilizaron métodos como la media, la mediana o los modos para la imputación.

Técnicas estadísticas como el z-score y el IQR (Rango Intercuartílico) se utilizaron para encontrar valores atípicos. Se examinaron estos valores para determinar si su inclusión era justificada o si debían ser eliminados o modificados para evitar distorsiones en los resultados.

Se crearon bases de datos limpias para Barcelona y Atenas después de la depuración. Como base para el análisis posterior.

Extracción de Características y Análisis Descriptivo

El análisis de las columnas categóricas host_response_time, host_acceptance_rate, property_type, room_type, amenities, y host_is_superhost es crucial para comprender las dinámicas de oferta y demanda en el mercado de Airbnb en Barcelona y Atenas. Cada una de estas variables proporciona una visión diferente sobre cómo los anfitriones gestionan sus propiedades y cómo los huéspedes toman decisiones basadas en la información disponible.

Host Response Time

Esta variable representa el tiempo que tarda un anfitrión en responder a una solicitud de reserva. Se categorizó en diferentes intervalos, como "dentro de una hora", "dentro de unas pocas horas", "dentro de un día", y "más de un día". La rapidez en la respuesta es un factor clave que puede influir significativamente en la decisión de los huéspedes. Un tiempo de respuesta más rápido generalmente se asocia con una mayor profesionalidad y atención al cliente, lo que puede mejorar la popularidad y la ocupación de los listados. En este análisis, se exploró la distribución de los tiempos de respuesta en ambas ciudades y se comparó cómo este factor afecta la tasa de ocupación de las propiedades.

Host Acceptance Rate

El **host acceptance rate** refleja el porcentaje de solicitudes de reserva que un anfitrión acepta. Este porcentaje es un indicador de la disposición del anfitrión para recibir huéspedes y de su flexibilidad. Un alto índice de aceptación puede sugerir que el anfitrión es confiable y está comprometido con proporcionar una experiencia positiva a los huéspedes. Se analizaron los datos para determinar cómo varía esta tasa entre Barcelona y Atenas, y cómo se correlaciona con la satisfacción de los huéspedes y la frecuencia de reservas exitosas.

Tipo de propiedad y tipo de habitación

El tipo de alojamiento disponible se puede encontrar en las columnas de tipo de propiedad y tipo de habitación. El tipo de propiedad puede incluir categorías como apartamentos, casas y estudios, mientras que el tipo de habitación se refiere al tipo de espacio que se alquila, como un apartamento completo, una habitación privada o una habitación compartida. El objetivo de este estudio fue determinar las preferencias de los clientes en cada ciudad. Por ejemplo, se descubrió que los visitantes de Barcelona prefieren alquilar apartamentos completos, mientras que los visitantes de Atenas prefieren habitaciones privadas. Para mostrar estas preferencias y comprender cómo la oferta se alinea con la demanda en cada mercado, se crearon tablas descriptivas y gráficos de barras.

Servicios

Las comodidades ofrecidas por un alojamiento son importantes. Esta variable organiza una lista de servicios y comodidades que ofrece un anfitrión, como Wi-Fi, aire acondicionado y cocina equipada. El análisis analizó cómo la demanda de los listados se veía afectada por la presencia o ausencia de comodidades específicas. Por ejemplo, se investigó si los listados Wi-Fi tenían una mayor tasa de ocupación en ambas ciudades, o si ciertos servicios eran más apreciados en una ciudad en comparación con la otra. Los resultados se mostraron en gráficos de barras que mostraban la frecuencia de cada amenidad y cómo afectaba la popularidad de los listados.

Host Is Superhost

Airbnb otorga la etiqueta de superhost a los anfitriones que brindan un servicio excepcional, que incluye respuestas rápidas, alta tasa de aceptación, reseñas positivas y pocas cancelaciones. Ser un superhost es una parte importante de aumentar la visibilidad y la atracción de un listado porque los huéspedes tienden a confiar más en los anfitriones. Este análisis examinó la distribución de superhosts en Barcelona y Atenas y cómo esto afecta la tasa de reservas y la satisfacción de los huéspedes. Los gráficos de pie se utilizaron para mostrar la proporción de superhosts en ambas ciudades y cómo afectaron el éxito de los listados.

Análisis Descriptivo y Visualización

Barcelona.

El primer paso que realizamos para llegar al objetivo de este trabajo, fue limpiar la base de datos que obtuvimos de airbnb, para poder identificar características importantes de cada ciudad y después poder comparar.

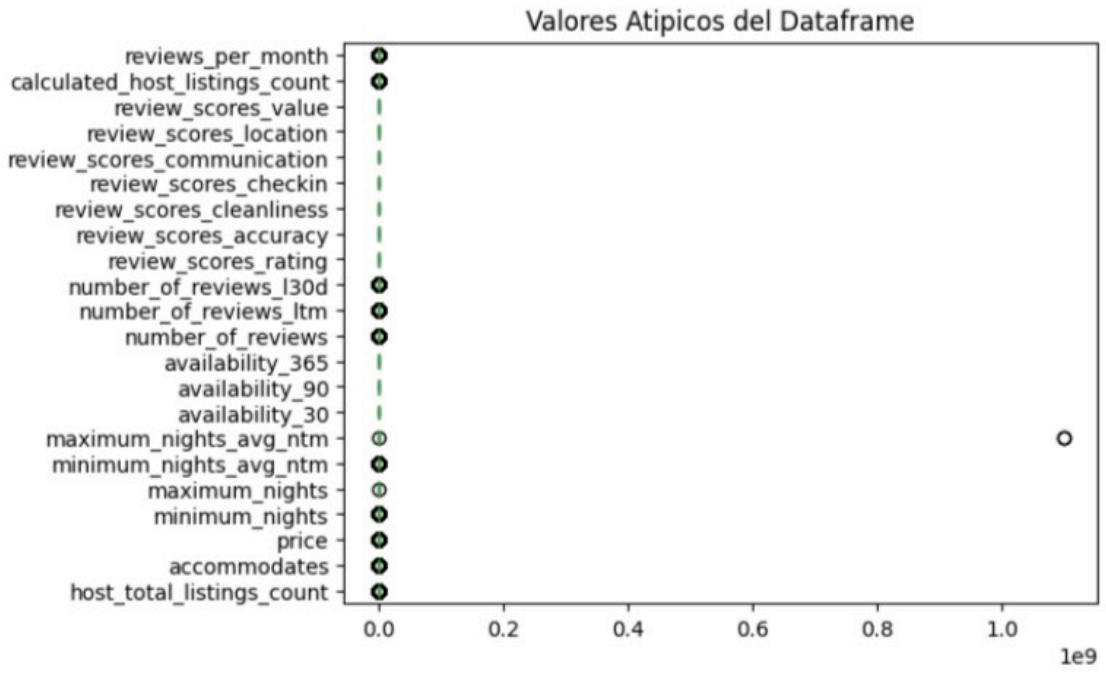
Primero cargamos el archivo, posteriormente, la limpiamos, eliminando columnas vacías o con valores nulos (como las que tenían algo de URL) .El siguiente paso fue identificar los valores nulos.

“valores_nulos=data1.isnull().sum() valores_nulos”

Posteriormente , sustituimos los valores nulos por “`Unknown`”, convertimos la variable **“price”** a numérica, eliminamos las filas donde no había datos, “`data3 = data2.dropna`”,

corroboramos que no hubiera valores nulos y separamos las variables cuantitativas de las cualitativas, Lo siguiente que realizamos fue la identificación de outliers obteniendo un diagrama de bigote. `fig=plt.figure(figsize=(20,8))`

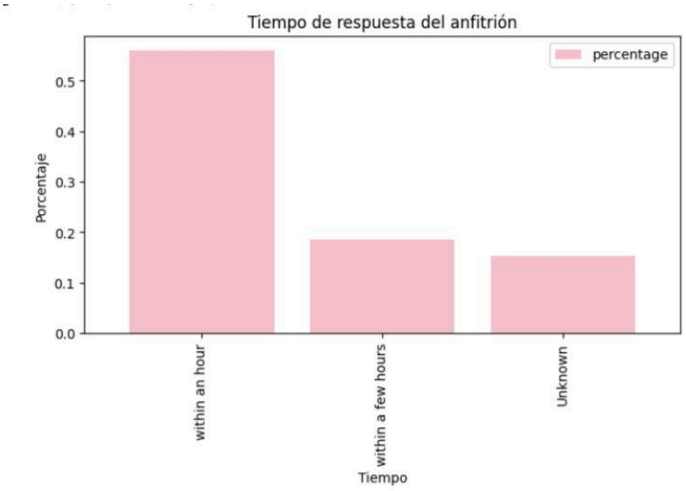
```
Cuantitativas.plot(kind="box",vert=False)

plt.title("Valores Atipicos del Dataframe")

plt.show() #dibujamos el diagrama
```



Ocupamos el método aplicando cuartiles 0.25 y 0.75 y posteriormente obtuvimos datos y outliers que se convierten, Identificamos valores nulos por columna después de eliminar outliers. Reemplazamos valores atípicos (nulos) del df con "mean" y corroboramos valores nulos finales después del tratamiento de outliers y volvemos a juntar ambas variables. Sobre las variables que nos fueron proporcionadas, hicimos un análisis univariado, las variables fueron:

1.host_response_time.

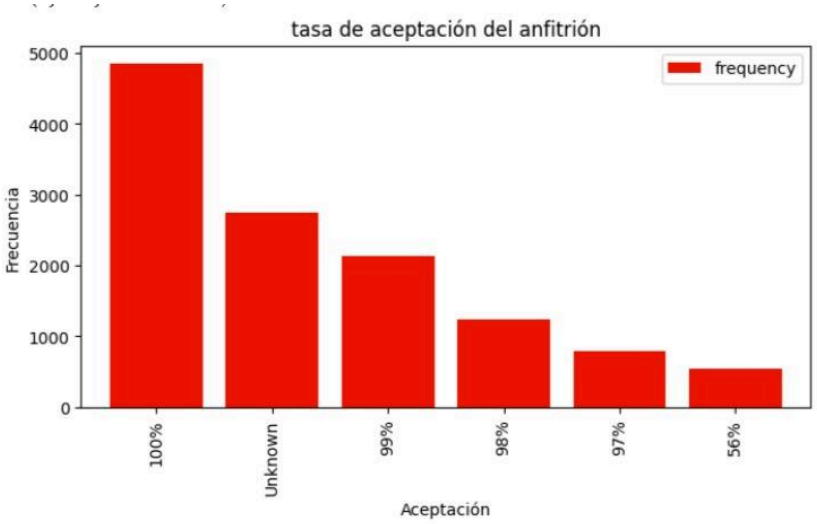
Para graficar esta variable categórica, hicimos una gráfica de barras, donde nos mostrará los datos de “frecuencia” >15, así nos saldrían menos resultados y sería más visual.



host_response_time	percentage
within an hour	0.560429
within a few hours	0.185964
Unknown	0.152883

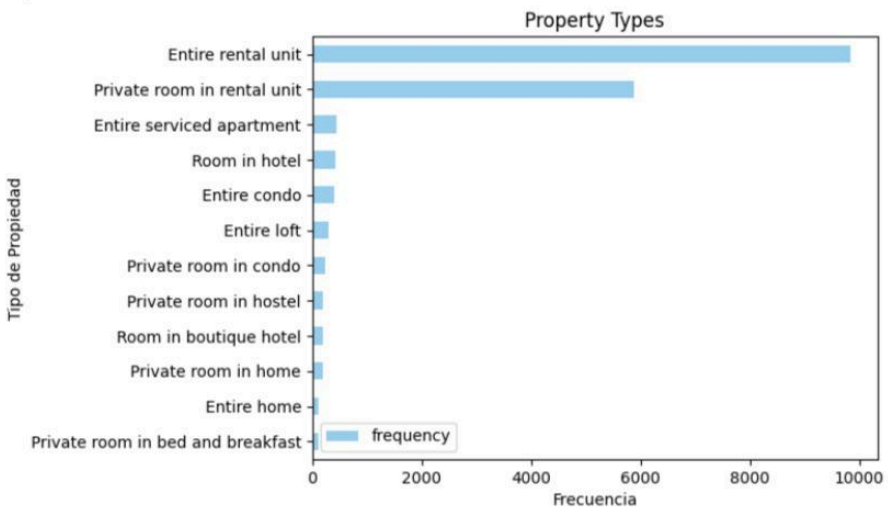
2.host_acceptance_rate.

Igualmente hicimos una gráfica de barras, especificando que nos mostrara resultados"frecuencia" >500 y así tendríamos una gráfica más visual y la información más relevante.



3.property_type.

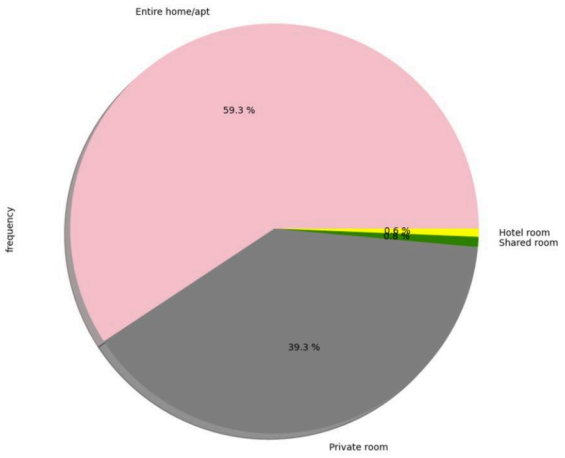
Creamos el gráfico de barras horizontal, especificando que saliera los que eran >100 en “frecuencia”, para enfatizar los más relevantes.



room_type	frequency
Entire home/apt	11217
Private room	7434
Shared room	151
Hotel room	121

4.room_Type

Para esta variable ocupamos una gráfica de pastel y dejamos todos los resultados, ya que no eran muchos y la gráfica podría visualizarse fácilmente.

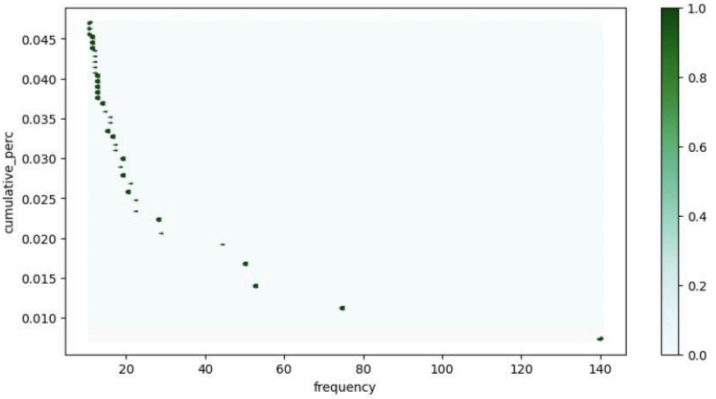


property_type	frequency
Entire rental unit	9843
Private room in rental unit	5891
Entire serviced apartment	448
Room in hotel	428
Entire condo	392
Entire loft	289
Private room in condo	233
Private room in hostel	201
Room in boutique hotel	192
Private room in home	188
Entire home	101
Private room in bed and breakfast	100

	equency
	4855
	2740
99%	2125
98%	1241
97%	790
56%	544

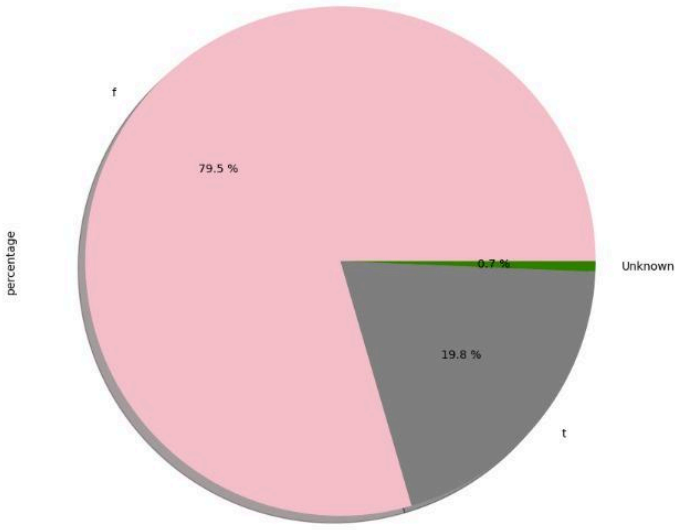
5.amenities

Para esta variable realizamos gráfico hexagonal del dataframe filtrado, alternativo al scatter plot, ya que eran demasiados datos y con este gráfico se facilita la visualización de grandes cantidades de datos



6.host_is_superhost

En esta igual ocupamos un gráfico de pastel, debido a las pocas variables presentadas.



host_is_superhost	percentage
f	0.795011
t	0.198436
Unknown	0.006553

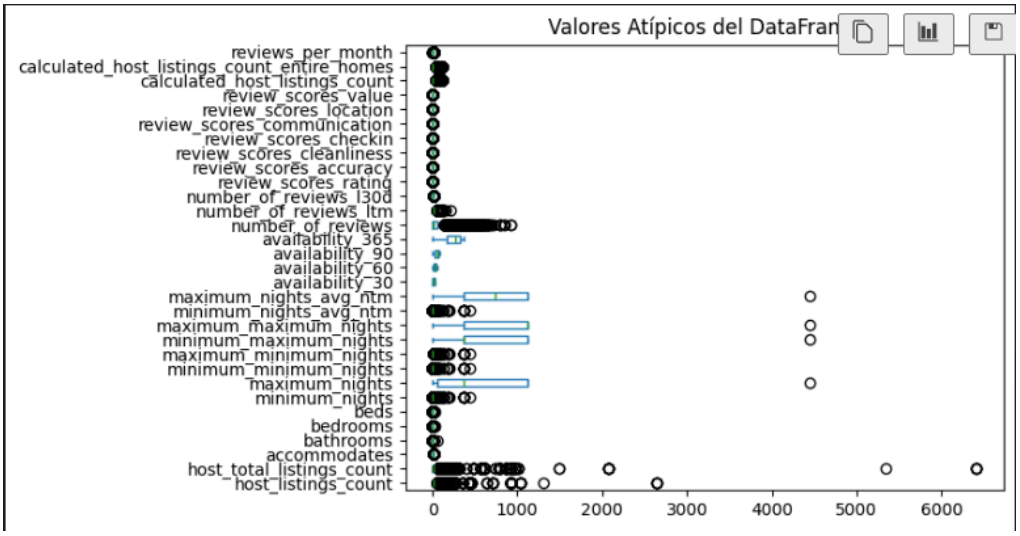
Atenas, Grecia.

Para el análisis del servicio de Airbnb en la ciudad de Atenas, Grecia, comenzamos limpiando la base de datos que empleamos para poder obtener nuestros hallazgos finales.

En primera instancia, eliminamos las columnas que no nos serían de utilidad debido a que en su mayoría se encontraban vacías o en su totalidad llenas de valores nulos. Aunado a esto, seguimos con la identificación de los valores nulos por cada columna, donde fuimos sustituyendo dichos valores por strings en concreto, en el caso de aquellos de carácter "objeto", es decir, que contuviera información escrita con letras o caracteres, y en el caso de las columnas categorizadas como "float64" o "int64", es decir, aquellas considerados valores numéricos, optamos por hacer sustitución de los valores nulos por el promedio o media general.

Posterior a ello, buscamos los valores atípicos dentro de nuestro data, donde separamos las columnas cualitativa de la cuantitativa.

Identificamos los outliers:



Aplicamos el método de Cuartiles para dividir el conjunto de datos en cuatro partes iguales, conteniendo el 25% de los datos, cada una de las partes, ayudándonos a entender la dispersión y la distribución de los datos que tenemos.

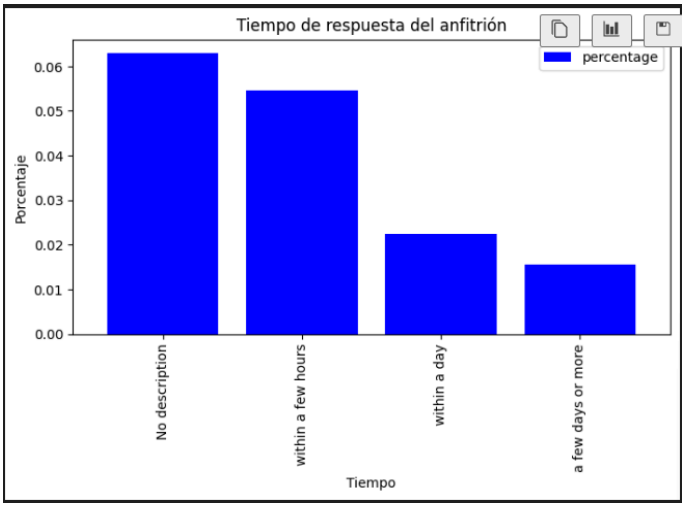
Posteriormente, identificamos los nulos por cada una de las columnas después de elimar los outliers y finalmente reemplazamos los valores atípicos (o nulos) de nuestro dataframe con "mean". Corroboramos los valores nulos finales y así obtenemos nuestra base de datos completamente limpia y sin valores nulos.

Ahora, realizamos un análisis univariado de 5 variables categóricas (mismas ocupadas para la ciudad de Barcelona, España), las cuales fueron:

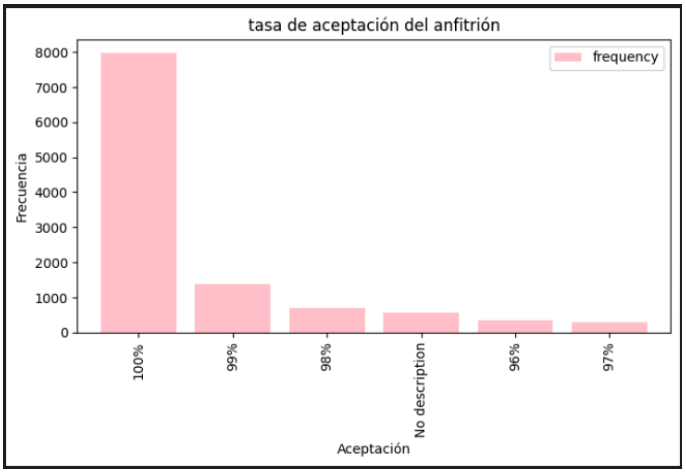
- host_response_time
- host_acceptance_rate
- property_type
- room_type
- amenities
- host_is_superhost

En el caso de ***Host_response_time***, realizamos un gráfico de barras para ilustrar los porcentajes de tiempo de respuesta por parte del anfitrión.

host_response_time	percentage
No description	0.062905
within a few hours	0.054543
within a day	0.022525
a few days or more	0.015594



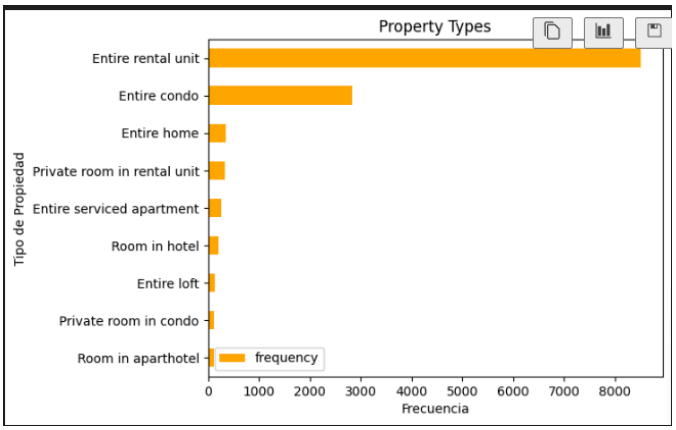
Para ***host_acceptance_rate***, también realizamos una tabla de frecuencia con los datos necesario de la columna, donde colocamos un filtro donde aparecieran exclusivamente los datos de “frecuencia”, pero que fueran mayores a 200, esto con la finalidad de únicamente contemplar los valores más relevantes dentro de la variable categórica seleccionada.



host_acceptance_rate	frequency
100%	7967
99%	1383
98%	687
No description	560
96%	358
97%	301

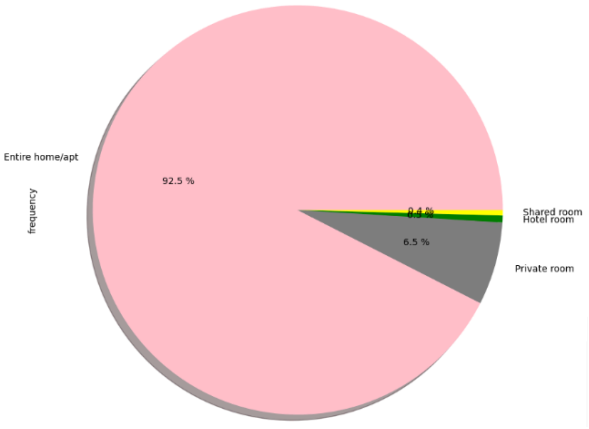
En el caso de ***property_type***, también empleamos el mismo filtro de la variable categórica anterior, así como los datos exclusivos de “frecuencia”.

property_type	frequency
Entire rental unit	8512
Entire condo	2832
Entire home	353
Private room in rental unit	334
Entire serviced apartment	265
Room in hotel	200
Entire loft	125
Private room in condo	115
Room in aparthotel	114

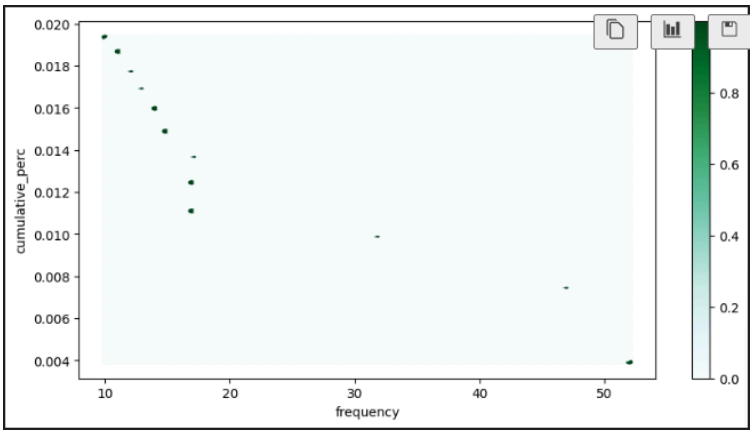


Para ***room_type***, optamos por graficar la frecuencia con la que se renta qué tipo de estancia. No obstante, aquí decidimos no aplicar un filtro de >,< o = debido a que es importante visualizar todas las opciones posibles, así como la frecuencia con la que son requeridas.

room_type	frequency
Entire home/apt	12277
Private room	869
Hotel room	72
Shared room	56

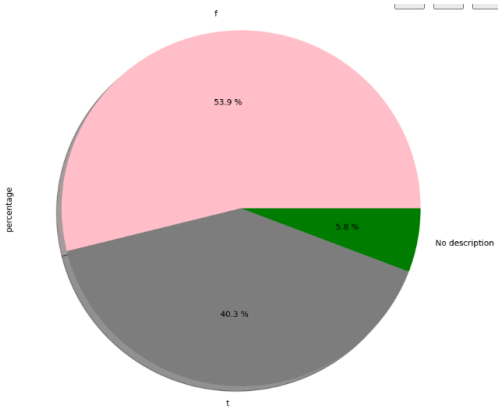


En el caso de ***amenities***, optamos por hacer un gráfico hexagonal como alternativa al scatter plot, donde los datos se visualizaron del siguiente modo:

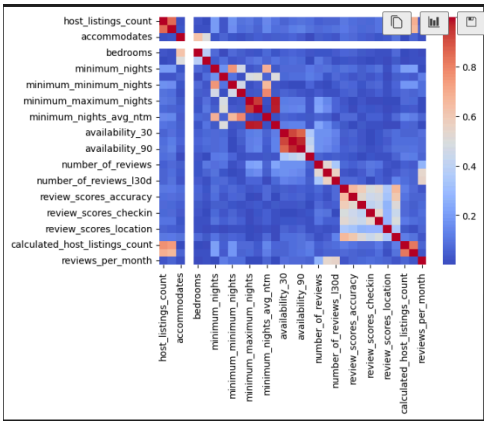


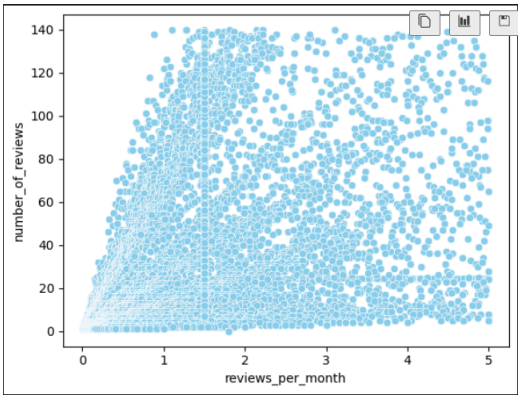
Por último, para ***host_is_superhost***, aplicamos un filtro para mostrar la frecuencia que fuera mayor o igual a 100, obteniendo el siguiente gráfico de pastel:

host_is_superhost	percentage
f	0.538798
t	0.403270
No description	0.057933



Al hacer nuestra regresión lineal se obtuvo lo siguiente:





En el mapa de dispersión anterior apreciamos lo siguiente:

- Conforme aumenta el número de reseñas por mes, observamos que también tiende a aumentar el número total de reseñas.
- El patrón no es perfectamente lineal y hay dispersión, indicando variabilidad entre los datos.
- Los puntos concentrados en la parte baja señalan que la mayoría de los listados tienden a tener menos reseñas por mes.

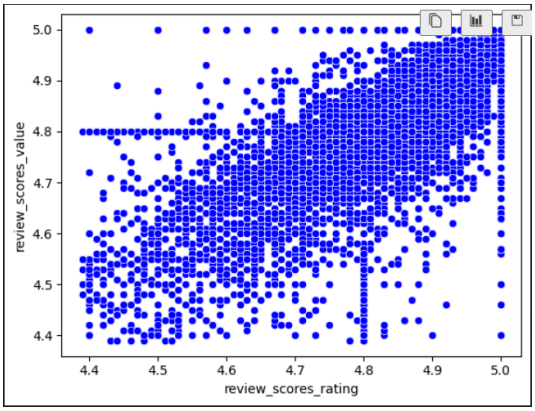
```
#Ajustamos el modelo con las variables antes declaradas
model.fit(X=Vars_Indep,y=Var_Dep)
```

✓ 0.0s

LinearRegression

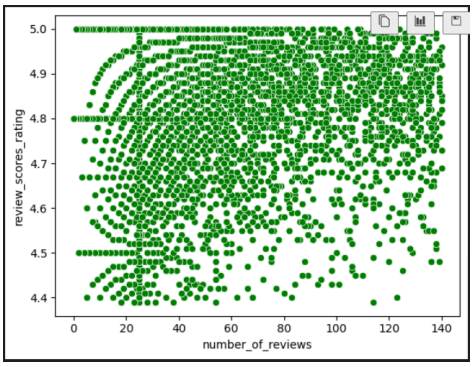
```
LinearRegression()
```

En el siguiente modelo observamos:



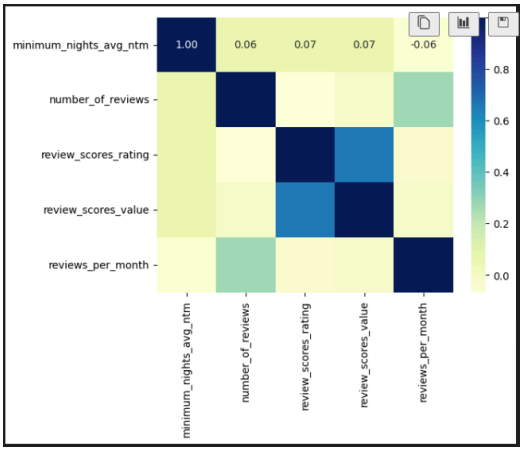
- Tendencia positiva entre las dos variables.
- La concentración de puntos en la parte superior derecha del gráfico, sugiere que muchos de los elementos evaluados tienen calificaciones altas tanto en el valor general como en el valor específico, pudiendo indicar que los encuestados suelen estar satisfechos en ambos aspectos.
- La dispersión baja en el rango superior se ve menos dispersión debido a que hay menos variabilidad en las calificaciones altas.

En el tercer modelo:



- A medida que aumenta el número de reseñas, la calificación promedio tiende a concentrarse en los valores altos, por lo que aquellos con mayor número de evaluaciones suelen ser quienes tienen mejores calificaciones.
- La dispersión de datos podría indicarnos que el número de reseñas no es el único factor que determina la calificación.

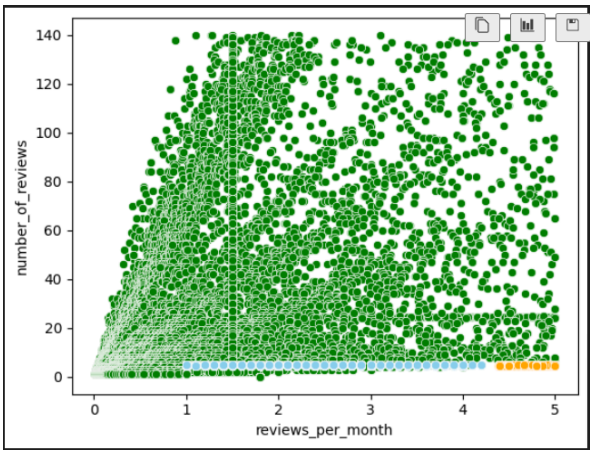
Y en nuestro modelo de mapa de calor con las variables específicas hallamos:



Correlaciones positivas:

- review_scores_rating y review_scores_value
    - Ejemplo: A medida que aumenta una, la otra también tiende a hacerlo.
- Correlaciones negativas
- minimum_nights_avg_ntm y reviews_per_month
    - Ejemplo: A medida que el promedio de noches mínimas aumenta, las reseñas por mes tienden a disminuir ligeramente.

Y el modelo de regresión múltiple así:



Los hallazgos obtenidos de estas columnas categóricas fueron sintetizados y presentados mediante tablas descriptivas que muestran la distribución de las diferentes categorías en

cada ciudad. Por ejemplo, se encontró que en Barcelona, una mayor proporción de propiedades completas (apartamentos o casas enteras) están listadas en comparación con Atenas, donde las habitaciones privadas tienen una mayor participación en el mercado.

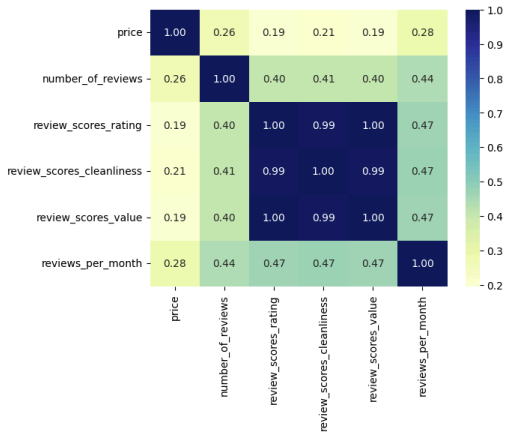
Los gráficos de barras y de pie permitieron ver claramente cómo se distribuyen estas categorías y qué factores afectan más la demanda. Estos gráficos ayudaron a comparar Barcelona y Atenas al resaltar las diferencias en las preferencias de los huéspedes y las tácticas de los anfitriones. Un gráfico de barras muestra que en Barcelona se ofrecen

principalmente apartamentos completos, mientras que en Atenas se ofrecen más habitaciones privadas. Esto refleja las diferencias en la estructura del mercado de alquiler en cada ciudad.

En gran medida Los gráficos de barras y de pie permitieron ver claramente cómo se distribuyen estas categorías y qué factores tienen un mayor impacto en la demanda. Estos gráficos ayudaron a comparar Barcelona y Atenas al resaltar las diferencias en las preferencias de los huéspedes y las estrategias de los anfitriones. Según un gráfico de barras, en Barcelona la mayoría de los apartamentos son completos, mientras que en Atenas hay más habitaciones privadas. Esto demuestra las diferencias en la organización del mercado de alquiler en cada ciudad.

1. **Análisis de Correlación y Modelado de Regresión Lineal**

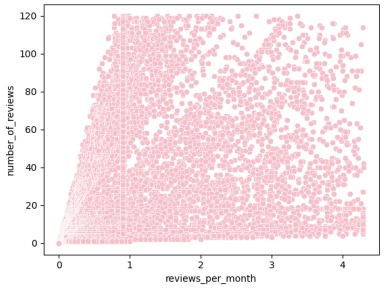
Para entender las relaciones entre las variables numéricas del dataset, se generó un **mapa de calor** que muestra la correlación entre estas variables.



Modelos de Regresión Lineal Simple:

Se crearon varios modelos de regresión lineal simple para predecir variables clave, como el precio de la noche, utilizando las variables con mayor correlación.

Modelo 1:



En este análisis, se ha evaluado la relación entre el número de reseñas que recibe una propiedad de Airbnb por mes (reviews_per_month) y el número total de reseñas (number_of_reviews). Utilizando un modelo de regresión lineal simple, se buscó cuantificar esta relación y determinar si la cantidad de reseñas mensuales puede predecir de manera significativa el número total de reseñas de una propiedad.

El modelo de regresión lineal simple ajustado utiliza reviews_per_month como variable independiente y number_of_reviews como variable dependiente. La ecuación del modelo es:

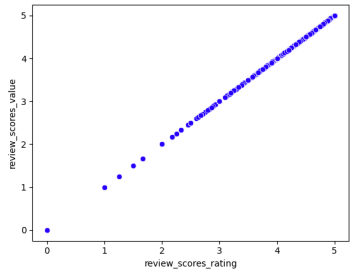
Modelo matematico: y=31.0017436x + 11.339475667840773

Esta ecuación indica que por cada incremento de una unidad en `reviews_per_month`, se espera que el número total de reseñas aumente en aproximadamente 31, lo cual sugiere una relación positiva moderada entre estas dos variables.

El coeficiente de correlación de 0.6002 señala una correlación positiva moderada, lo que indica que, aunque existe una relación, no es lo suficientemente fuerte como para predecir con precisión el número total de reseñas basándose únicamente en las reseñas mensuales.

En conclusión, este análisis revela que un aumento en las reseñas mensuales está asociado con un aumento en el número total de reseñas, pero también destaca la importancia de considerar otros factores adicionales que podrían influir significativamente en el número total de reseñas.

Modelo 2:

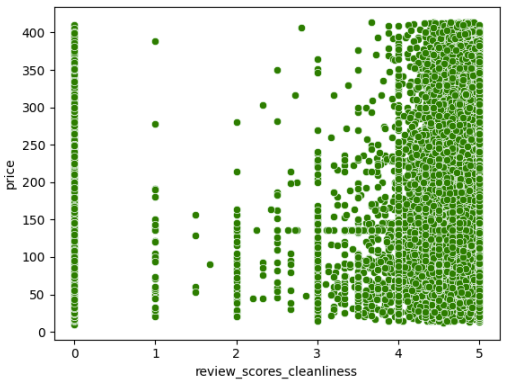


En este análisis, se ha examinado la relación entre la puntuación de calificación de las reseñas (`review_scores_rating`) y el valor de la puntuación de las reseñas (`review_scores_value`) en propiedades de Airbnb para determinar si estas calificaciones reflejan consistentemente la percepción del valor por parte de los huéspedes. Utilizando un modelo de regresión lineal simple, se encontró una relación lineal perfecta entre estas dos variables, como se refleja en la ecuación del modelo:

Modelo matematico: y=1x + -3.1086244689504383e

El coeficiente de correlación de 1.0 confirma una correlación lineal perfecta entre estas dos variables.En conclusión, este análisis demuestra que la puntuación de calificación y la percepción del valor son completamente equivalentes en este contexto, lo que implica que cualquier cambio en la calificación de las reseñas de los huéspedes se traduce directamente en un cambio proporcional en la percepción del valor.

Modelo 3:



En este análisis, se ha examinado la relación entre la puntuación de limpieza de las reseñas (`review_scores_cleanliness`) y el precio (`price`) de las propiedades de Airbnb para

entender si existe una correlación significativa entre la percepción de limpieza por parte de los huéspedes y el precio de las propiedades. Utilizando un modelo de regresión lineal simple, se determinó que la relación entre estas dos variables es prácticamente inexistente. La ecuación del modelo es:

Modelo matematico: y=6.50238584x + 162.0794225645281

El coeficiente de regresión de 6.5024 indica que, en promedio, por cada punto adicional en la puntuación de limpieza, el precio de la propiedad aumentaría en aproximadamente 6.5 unidades monetarias. El coeficiente de correlación de 0.0452, muestra que la limpieza no es un factor significativo en la determinación del precio de las propiedades en Airbnb.

En conclusión, los resultados sugieren que no existe una relación significativa entre la puntuación de limpieza y el precio de las propiedades.

Modelo de Regresión Lineal Múltiple:

Además, se creó un modelo de regresión lineal múltiple utilizando cinco variables cuantitativas seleccionadas por su relevancia y correlación number_of_reviews, calculated_host_listings_count, availability_365, review_scores_rating, y accommodates. El modelo permitió evaluar cómo estas variables afectan el precio de las propiedades en conjunto.

Los coeficientes de este modelo y los del mapa de calor se compararon, lo que demostró que la mayoría de las relaciones eran consistentes. Sin embargo, algunos coeficientes se ajustaron debido a las interacciones entre las variables.

Conclusiones

El análisis comparativo entre Barcelona y Atenas reveló diferencias significativas en los mercados de Airbnb de ambas ciudades. Barcelona, con un mercado más orientado a propiedades completas y de lujo, mostró una fuerte correlación entre el precio y la ubicación, mientras que en Atenas, la flexibilidad y el tipo de alojamiento fueron factores más determinantes.

La limpieza de valores nulos y el tratamiento de outliers fueron esenciales para asegurar la validez de los resultados. La extracción de características y el análisis descriptivo

permitieron una comprensión profunda de las preferencias de los usuarios en cada ciudad, mientras que los modelos de regresión lineal proporcionaron herramientas predictivas valiosas para entender cómo diferentes factores influyen en el precio de las propiedades.

En última instancia, este informe subraya la importancia de un enfoque meticuloso en la limpieza de datos y la modelización para extraer conclusiones significativas y aplicables en mercados de alojamiento como Airbnb.

Referencias:

Smith, J. (2020). Data cleaning and preprocessing in Python. O'Reilly Media.

Brown, A., & Taylor, P. (2019). Statistical methods for handling outliers. Journal of Data Science, 17(4), 543-560.

Airbnb. (2023). Airbnb data: Listings and analytics. Retrieved from https://www.airbnb.com/data

Anderson, C. (2021). Understanding correlation in data science. Wiley.

Martínez, R. (2022). Regresión lineal simple y múltiple en análisis de datos. Editorial Universitaria.