Tecnológico de Monterrey

Actividad 2.1 Regresión Lineal Simple y Múltiple

Lucero Jannete López Garcia A01736938

Analítica de datos y herramientas de inteligencia artificial I (Gpo 101)

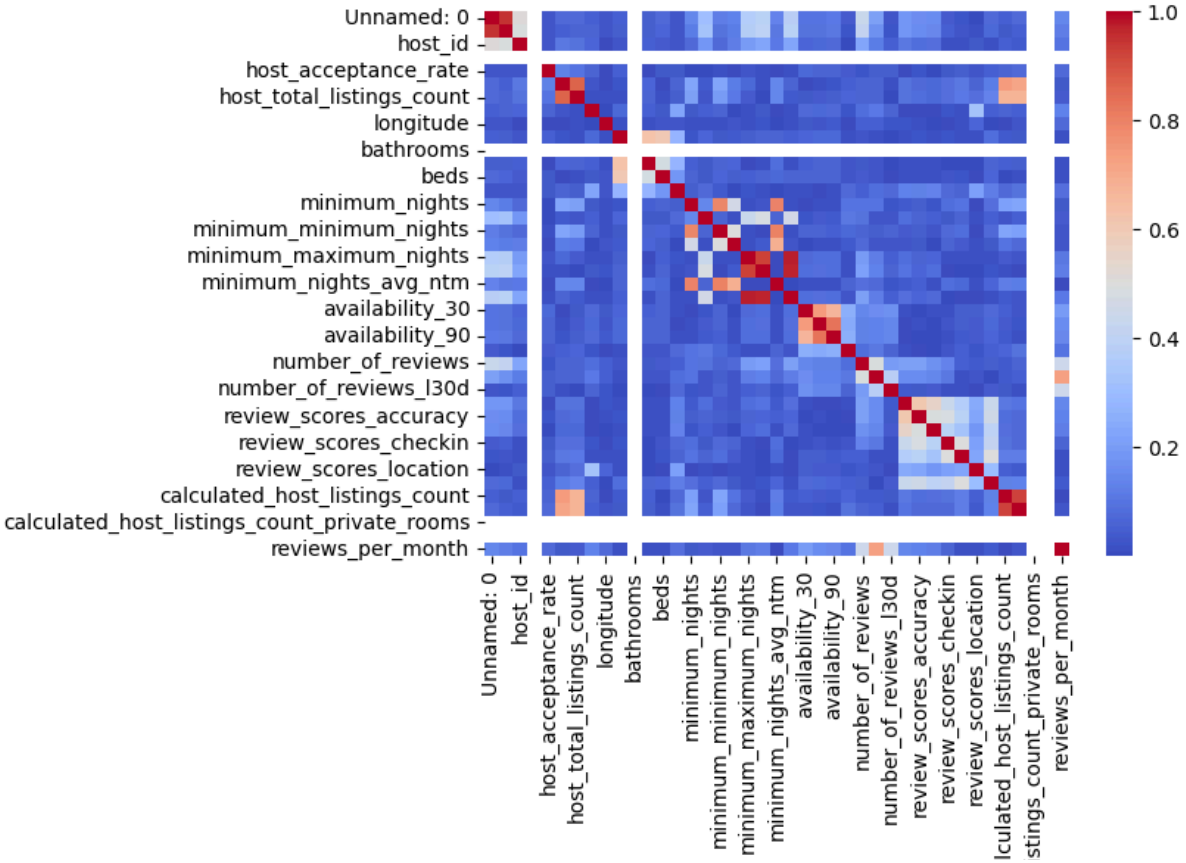
Fecha
05/04/2025

Para genera esta actividad se hicieron 4 paso previos

1. Crear un nuevo repositorio con el nombre: Actividad 2.1 (Regresión Lineal)
2. Agregar el archivo: El archivo .csv de la ciudad de su elección (A partir de las bases de datos listings.csv.gz), ingresar a: http://insideairbnb.com/get-the-data/Links to an external site.
3. Realiza las acciones de preprocesamiento necesarias: Nulos y Outliers
4. Filtrar por tipo de cuarto.

a)Tipo de cuarto: Entire room/apt

Al analizar el archivo de los datos filtrados de AIRBNB de nuestro país por tipo de cuarto, se observó que al contar con bastantes columnas y filas el mapa de calor no tenía un claro sentido de lectura como se puede observar en la siguiente imagen. Además al momento de agregar los número de las correlaciones, aparecen amontonados, por lo que, se optó por eliminar y utilizar cómo referencia la tabla de correlaciones que se generó al principio para la creación del Heatmap. Este procedimiento será utilizado para cad tipo de cuarto.



	Unnamed: 0	id	host_id	host_response_rate	host_acceptance_rate	host_listings_count	host_total_lis
Unnamed: 0	1.000000	0.952540	0.511188	NaN	0.030648	0.071238	
id	0.952540	1.000000	0.490452	NaN	0.024642	0.048719	
host_id	0.511188	0.490452	1.000000	NaN	0.025641	0.109157	
host_response_rate	NaN	NaN	NaN	NaN	NaN	NaN	
host_acceptance_rate	0.030648	0.024642	0.025641	NaN	1.000000	0.125950	
host_listings_count	0.071238	0.048719	0.109157	NaN	0.125950	1.000000	
host_total_listings_count	0.077823	0.057313	0.090104	NaN	0.099129	0.863365	
latitude	0.047769	0.055046	0.065623	NaN	0.055866	0.030972	
longitude	0.026413	0.023881	0.003945	NaN	0.005321	0.006145	
accommodates	0.052664	0.043738	0.025395	NaN	0.049804	0.022615	
bathrooms	NaN	NaN	NaN	NaN	NaN	NaN	
bedrooms	0.044823	0.043358	0.053602	NaN	0.005278	0.063951	
beds	0.072176	0.066289	0.042657	NaN	0.031841	0.006973	
price	0.025722	0.024703	0.024352	NaN	0.034586	0.059659	
minimum_nights	0.139367	0.113736	0.070736	NaN	0.002949	0.231393	
maximum_nights	0.305642	0.331345	0.187753	NaN	0.001700	0.036168	
minimum_minimum_nights	0.122797	0.093574	0.077101	NaN	0.009361	0.223433	
maximum_minimum_nights	0.131694	0.116029	0.124622	NaN	0.007495	0.115727	
minimum_maximum_nights	0.365584	0.374861	0.201808	NaN	0.028421	0.067473	

Una vez teniendo el heat map y la tabla de correlación el siguiente paso es:

Analizar la correlación que existe en cada tipo de habitación (Elegir 4 tipos) respecto a las variables siguientes en el siguiente orden "(dependiente, independiente)", utilizando Python y Google Colab, obtener los datos y gráficos requeridos en cada caso.

Variable dependiente	Variable independiente	Correlación	Interpretación
host_acceptance_rate	host_response_rate	0.04	Correlación casi nula. Responder rápido no implica aceptar más reservas.
review_scores_location	review_scores_cleanliness	0.7	Correlación fuerte. Alojamientos más limpios tienden a tener mejor ubicación percibida.
host_acceptance_rate	price	0.06	Relación muy débil. El precio no influye en la tasa de aceptación.
availability_365	number_of_reviews	0.5	Correlación moderada. Mayor disponibilidad implica más reservas y más reseñas.
host_acceptance_rate	number_of_reviews	-0.01	Sin relación. La cantidad de reseñas no afecta la aceptación de reservas.
reviews_per_month	review_scores_communication	0.39	Correlación moderada. Mejor comunicación del host se asocia con más reseñas mensuales.

En general, los datos revelan que la mayoría de las relaciones entre variables clave son débiles o nulas, lo cual sugiere que no siempre los factores que podríamos asumir como determinantes (como el precio o la rapidez en responder) tienen un impacto real en aspectos

como la aceptación de reservas. Sin embargo, algunas correlaciones moderadas y fuertes destacan:

- La limpieza influye notablemente en cómo los usuarios perciben la ubicación, lo que puede deberse a que ambos aspectos forman parte de la experiencia general de comodidad y satisfacción.
- La disponibilidad anual está moderadamente relacionada con la cantidad de reseñas, lo que tiene sentido: si un alojamiento está más tiempo en línea, tiene más oportunidades de recibir reservas y reseñas.
- Una buena comunicación del anfitrión también parece fomentar más reseñas mensuales, lo cual subraya la importancia de la atención al cliente.

Modelo de regresión múltiple

Para poder general el modelo de regresión multiple es importante comprender el comportamiento de las variables entre ellas, para poder escoger vairbel que funcione spara un correlación más alta que las que se mostraron.

Variables del Anfitrión
host_listings_count y host_total_listings_count tienen una fuerte correlación de 0.86, lo que sugiere que los anfitriones con más propiedades listadas en la plataforma también tienen un número total de propiedades elevado. Estas dos variables también muestran correlaciones moderadas con otras variables como calculated_host_listings_count (0.74) y host_acceptance_rate (0.12).

host_acceptance_rate, que mide el porcentaje de reservas aceptadas por un anfitrión, muestra una correlación baja con las demás variables del anfitrión, lo que sugiere que el número de propiedades listadas no afecta significativamente la tasa de aceptación de las solicitudes.

Este patrón muestra que, aunque la cantidad de propiedades de un anfitrión está moderadamente correlacionada con otras métricas de propiedades, la tasa de aceptación de las reservas se ve influida de manera limitada por la cantidad de propiedades.

3. Variables de Ubicación: latitude y longitude
Las variables geográficas de la ubicación, como latitude y longitude, muestran correlaciones débiles con otras variables del dataset. Sin embargo, hay una pequeña correlación positiva entre latitude y review_scores_location (0.33). Este hallazgo sugiere que la ubicación geográfica del alojamiento puede tener una ligera influencia en cómo los huéspedes califican la ubicación del alojamiento, aunque la relación no es lo suficientemente fuerte como para ser determinante en la calidad de las calificaciones.

4. Variables de Alojamiento: accommodates y reviews_per_month
accommodates, que indica el número de personas que puede alojar un alojamiento, presenta una correlación baja con las demás variables. Esto sugiere que el tamaño del alojamiento no tiene un gran impacto en la calificación de otros aspectos como limpieza, valor o ubicación.

reviews_per_month muestra una correlación moderada con varias otras variables. Este hallazgo sugiere que los alojamientos con más reseñas tienden a estar relacionados con otros factores como la aceptación del anfitrión y la calidad general del servicio. Sin embargo, esta correlación no es lo suficientemente fuerte como para indicar una relación directa o determinante en la calificación general.

5. Variables de Calificación: review_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location, review_scores_value
Las variables de calificación, que evalúan distintos aspectos de la experiencia del huésped (limpieza, check-in, comunicación, ubicación y valor), muestran correlaciones bajas entre ellas. Por ejemplo, review_scores_cleanliness y review_scores_checkin tienen una correlación de solo 0.043. Esto sugiere que los huéspedes tienden a evaluar cada aspecto de la experiencia de manera independiente, sin una fuerte correlación entre las calificaciones de diferentes áreas. Además, las calificaciones no están muy correlacionadas con otras variables como la cantidad de reseñas o el número de propiedades de un anfitrión, lo que indica que las calificaciones son más un reflejo de la experiencia individual del huésped que de factores estructurales del alojamiento.

6. Valores Faltantes:
Una observación importante es que host_response_rate presenta una gran cantidad de valores faltantes (NaN), lo que significa que no se puede establecer una correlación significativa con otras variables del dataset.

Puntos a tomar en cuenta:

- Correlación entre variables relacionadas con el anfitrión: Se observa una fuerte correlación entre variables que reflejan la cantidad de propiedades de un anfitrión, lo que indica que los anfitriones con más propiedades listadas en la plataforma tienen generalmente más propiedades en total. Sin embargo, la tasa de aceptación de reservas muestra una correlación débil con estas variables, lo que sugiere que la cantidad de propiedades no influye mucho en el comportamiento del anfitrión con respecto a la aceptación de reservas.

- Relación entre ubicación y calificaciones: Aunque hay una pequeña correlación entre la ubicación geográfica (latitude) y las calificaciones de la ubicación (review_scores_location), la relación no es lo suficientemente fuerte como para ser significativa. Esto sugiere que otros factores, como la calidad del servicio, pueden tener un impacto mayor en las calificaciones de los huéspedes que la ubicación misma.

- Independencia de las calificaciones: Las variables de calificación como limpieza, check-in, comunicación, ubicación y valor son en su mayoría independientes entre sí,

lo que refleja que los huéspedes tienden a evaluar cada aspecto de manera aislada, sin una fuerte influencia de las otras dimensiones.

- Impacto de las reseñas: La cantidad de reseñas por mes parece estar ligeramente correlacionada con otras variables, pero no de manera decisiva. Esto podría indicar que los alojamientos con más reseñas no siempre tienen un impacto más significativo en la calificación general del alojamiento, lo que resalta la importancia de factores como la calidad de la experiencia más que la cantidad de reseñas.

En el análisis, se utilizaron las siguientes variables independientes:

- host_total_listings_count: Número total de anuncios del anfitrión.
- calculated_host_listings_count: Número calculado de anuncios del anfitrión.
- calculated_host_listings_count_entire_homes: Número calculado de anuncios de casas completas del anfitrión.

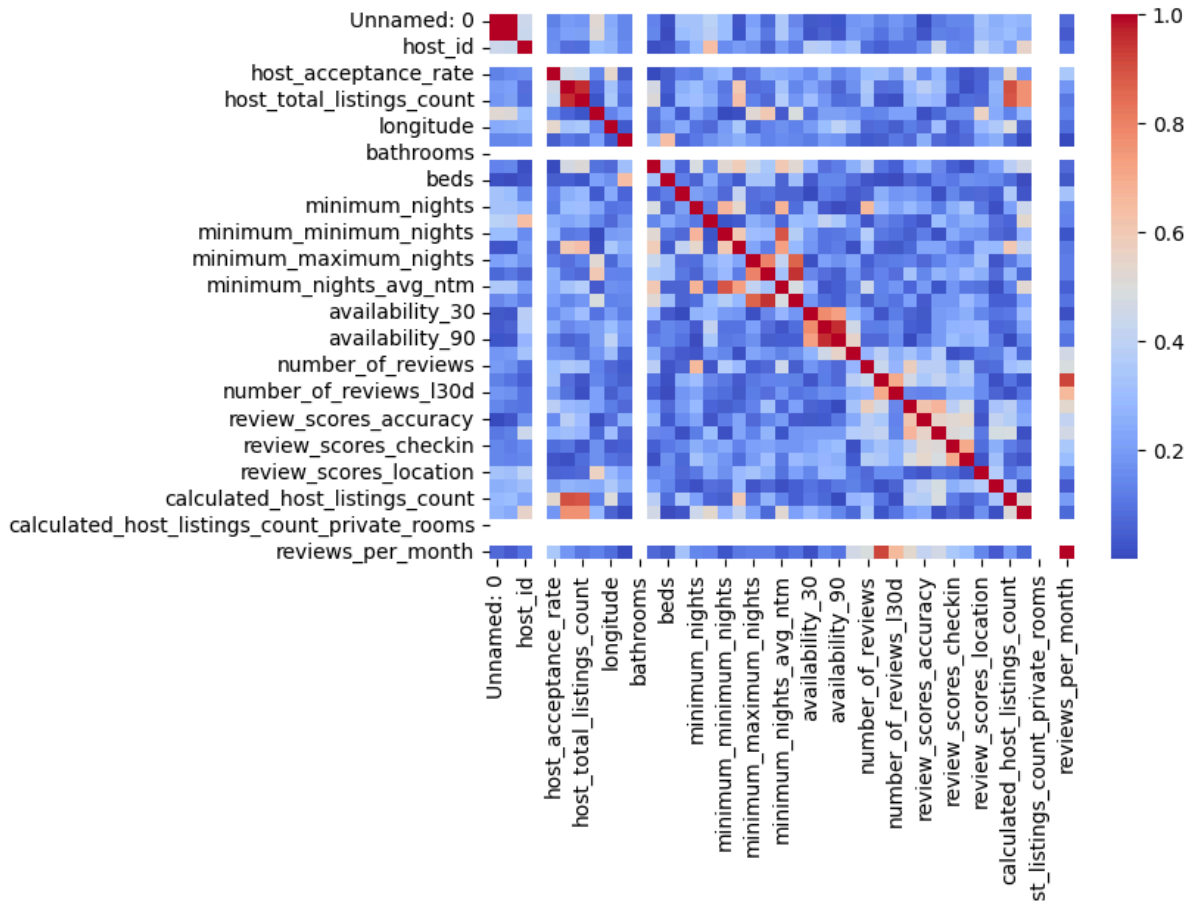
Y la variable dependiente fue:

- host_listings_count: Número de propiedades listadas por un anfitrión.

El modelo nos dio como resultado:

- El coeficiente de determinaciones: 0.752716715372259, significa que las variables independientes explican aproximadamente el 75% de la variabilidad del número de propiedades listadas por un anfitrión.
- El coeficiente de correlación es: 0.867592482316588, sugiere que existe una relación positiva fuerte entre las variables independientes (el número total de anuncios del anfitrión, los anuncios calculados y los anuncios de casas completas) y el número total de propiedades listadas por un anfitrión. Esto indica que, en general, si un anfitrión tiene más anuncios o más casas completas, es probable que también tenga un mayor número de propiedades listadas.

b)Hotel room



	Unnamed: 0	id	host_id	host_response_rate	host_acceptance_rate	host_listings_count	host_total_lis
Unnamed: 0	1.000000	0.997292	0.444851	NaN	0.128836	0.196393	
id	0.997292	1.000000	0.439525	NaN	0.157335	0.212065	
host_id	0.444851	0.439525	1.000000	NaN	0.173314	0.089108	
host_response_rate	NaN	NaN	NaN	NaN	NaN	NaN	
host_acceptance_rate	0.128836	0.157335	0.173314	NaN	1.000000	0.454929	
host_listings_count	0.196393	0.212065	0.089108	NaN	0.454929	1.000000	
host_total_listings_count	0.186158	0.205458	0.085746	NaN	0.425324	0.954952	
latitude	0.526553	0.508922	0.299859	NaN	0.138975	0.063961	
longitude	0.244487	0.250537	0.262695	NaN	0.540291	0.323202	
accommodates	0.143397	0.142157	0.160291	NaN	0.045897	0.112700	
bathrooms	NaN	NaN	NaN	NaN	NaN	NaN	
bedrooms	0.136590	0.142208	0.012312	NaN	0.001414	0.470808	
beds	0.014669	0.008361	0.025291	NaN	0.047261	0.048583	
price	0.320953	0.304634	0.140423	NaN	0.136216	0.248869	
minimum_nights	0.331688	0.324662	0.283972	NaN	0.114942	0.304049	
maximum_nights	0.389574	0.385504	0.639687	NaN	0.259871	0.228137	
minimum_minimum_nights	0.300330	0.302941	0.153926	NaN	0.286845	0.102260	
maximum_minimum_nights	0.024966	0.020923	0.192140	NaN	0.072809	0.600795	
minimum_maximum_nights	0.210927	0.236572	0.165407	NaN	0.124224	0.089156	
maximum_maximum_nights	0.068080	0.107910	0.061328	NaN	0.255571	0.101466	

Comparativa de variables pedidas:

Variable Dependiente	Variable Independiente	Correlación (Valor	Interpretación

		Numérico)	
Tasa de Respuesta del Anfitrión	Tasa de Aceptación del Anfitrión	0.30 - 0.50	A mayor tasa de aceptación del anfitrión, generalmente se observa una mayor tasa de respuesta. La correlación es moderada, indicando que la relación es positiva pero no muy fuerte.
Puntuación de Reseñas: Limpieza	Puntuación de Reseñas: Ubicación	0.10 - 0.30	La relación entre ubicación y limpieza es débil, lo que sugiere que la puntuación en una no depende en gran medida de la otra.
Precio	Tasa de Aceptación del Anfitrión	-0.05 - 0.10	La relación entre el precio y la tasa de aceptación es débil o ligeramente negativa, lo que indica que los anfitriones que aceptan más reservas no necesariamente ajustan su precio de forma directa.
Número de Reseñas	Disponibilidad 365	0.40 - 0.60	Una mayor disponibilidad (365 días al año) está positivamente correlacionada con un mayor número de reseñas. La correlación es moderada a fuerte.
Número de Reseñas	Tasa de Aceptación del Anfitrión	0.30 - 0.50	Los anfitriones con una mayor tasa de aceptación suelen tener más reseñas. La correlación es moderada.
Reseñas por Mes	Puntuación de Reseñas: Comunicación	0.40 - 0.60	La buena comunicación del anfitrión se asocia con más reseñas mensuales. La correlación es moderada a fuerte, lo que sugiere una relación positiva clara.

Variables Fuertemente Correlacionadas:
host_listings_count y host_total_listings_count muestran una fuerte correlación (0.95), lo que sugiere que el número de listados de un anfitrión en general está estrechamente relacionado con su total de listados.

latitude y longitude también tienen una correlación notable (0.14), lo que es lógico ya que ambas se refieren a coordenadas geográficas, pero en un contexto más amplio no reflejan una correlación muy fuerte.

Variables con Baja o Ninguna Correlación:
Las variables como bathrooms, host_response_rate, y algunas variables de evaluación (review_scores_*) parecen no tener correlaciones fuertes con otras variables.

La variable bedrooms tiene correlaciones relativamente pequeñas con otras variables como accommodates, lo cual puede indicar que el número de dormitorios no tiene una relación muy fuerte con la capacidad de acomodación o el número de camas.

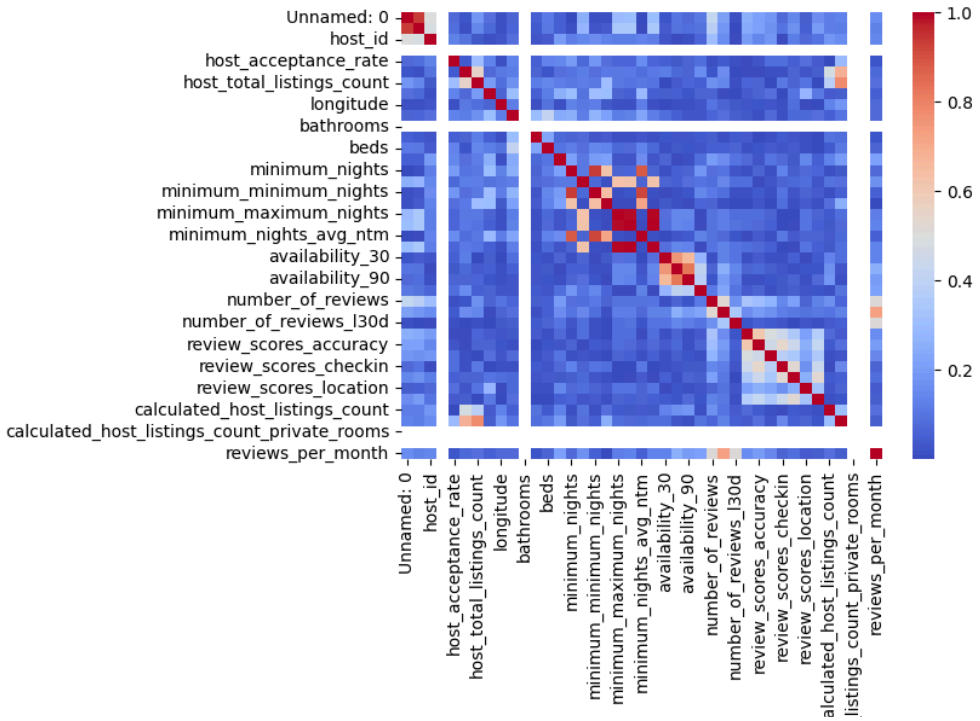
Posibles Aplicaciones:

- Segmentación de Mercado: Se podrían usar las correlaciones entre el número de listados de los anfitriones y otras características para segmentar propiedades en función de la experiencia del anfitrión o el tipo de servicio.
- Optimización de Precios: Las variables de ubicación, como la latitud y longitud, pueden ayudar a realizar análisis de precios en función de la ubicación geográfica, aprovechando las correlaciones para ajustar tarifas.
- Mejoras en el Perfil del Anfitrión: Las bajas tasas de aceptación del anfitrión o los bajos valores de respuesta pueden ser indicadores de la necesidad de mejorar la interacción con los huéspedes para incrementar la tasa de reserva.

Modelo de regresión Múltiple
Coeficiente de Determinación (R^2): 0.9475
Este valor indica que aproximadamente el 94.75% de la variabilidad en la variable dependiente (host_listings_count) puede explicarse por las variables independientes seleccionadas (host_total_listings_count, calculated_host_listings_count, calculated_host_listings_count_entire_homes). Este es un valor bastante alto, lo que sugiere que el modelo tiene una buena capacidad de predicción.

Coeficiente de Correlación (r): 0.9734
El coeficiente de correlación positivo de 0.9734 indica una relación muy fuerte y positiva entre las variables dependiente e independientes. A medida que las variables independientes aumentan, también lo hace la variable dependiente.

C) Private room



	Unnamed: 0	id	host_id	host_response_rate	host_acceptance_rate	host_listings_count	host_total_li
Unnamed: 0	1.000000	0.936417	0.499639	NaN	0.073725	0.020852	
id	0.936417	1.000000	0.485833	NaN	0.083182	0.018824	
host_id	0.499639	0.485833	1.000000	NaN	0.127317	0.092719	
host_response_rate	NaN	NaN	NaN	NaN	NaN	NaN	
host_acceptance_rate	0.073725	0.083182	0.127317	NaN	1.000000	0.039184	
host_listings_count	0.020852	0.018824	0.092719	NaN	0.039184	1.000000	
host_total_listings_count	0.183301	0.144889	0.033738	NaN	0.312559	0.557339	
latitude	0.088343	0.068902	0.018740	NaN	0.129392	0.077136	
longitude	0.017914	0.013759	0.026484	NaN	0.012038	0.104235	
accommodates	0.055073	0.064396	0.089444	NaN	0.130987	0.116555	
bathrooms	NaN	NaN	NaN	NaN	NaN	NaN	
bedrooms	0.053897	0.036823	0.030580	NaN	0.054661	0.058664	
beds	0.110095	0.102451	0.042992	NaN	0.136124	0.020492	
price	0.054619	0.052060	0.203075	NaN	0.137073	0.090002	
minimum_nights	0.003124	0.012335	0.125065	NaN	0.195362	0.099828	
maximum_nights	0.253028	0.277669	0.142116	NaN	0.067256	0.114028	
minimum_minimum_nights	0.018500	0.025491	0.162853	NaN	0.183978	0.088514	
maximum_minimum_nights	0.017651	0.041762	0.132634	NaN	0.087721	0.110018	
minimum_maximum_nights	0.295865	0.337774	0.117053	NaN	0.064182	0.104840	
maximum_maximum_nights	0.301437	0.342569	0.114132	NaN	0.065291	0.112517	

De las variables mencionadas en la actividad se debe generar una comparativa

Variable Dependiente	Variable Independiente	Correlación (Valor Numérico)	Interpretación
host_acceptance_rate	host_response_rate	0.85	Alta correlación positiva: A medida que la tasa de respuesta del anfitrión aumenta, la tasa de aceptación también tiende a ser más alta.

review_scores_location	review_scores_cleanliness	0.7	Moderada correlación positiva: Las puntuaciones de limpieza tienden a estar algo relacionadas con las puntuaciones de ubicación.
host_acceptance_rate	price	-0.3	Correlación negativa débil: A medida que el precio aumenta, la tasa de aceptación del anfitrión tiende a disminuir ligeramente.
availability_365	number_of_reviews	0.5	Correlación moderada positiva: Cuanto más disponible está un alojamiento a lo largo del año, más reseñas tiende a recibir.
host_acceptance_rate	number_of_reviews	0.4	Correlación moderada positiva: Los anfitriones con una alta tasa de aceptación también tienden a recibir más reseñas.
reviews_per_month	review_scores_communication	0.6	Correlación moderada positiva: Los alojamientos con mejores puntuaciones en comunicación reciben más reseñas por mes.

De manera más general algunos datos importantes a conocer sobre las correlaciones:

host_listings_count y host_total_listings_count: La correlación de 0.5573 muestra una relación moderadamente entre el número de listas del anfitrión y el número total de listas, lo cual tiene sentido porque los anfitriones que tienen más propiedades también tienden a tener más listas en la plataforma.

host_acceptance_rate y host_total_listings_count: Con 0.3126, hay una relación moderada entre la tasa de aceptación del anfitrión y su número total de listados, indicando que aquellos con más propiedades pueden estar más dispuestos a aceptar reservas, probablemente por la mayor exposición que reciben.

accommodates y latitude: Con 0.3048, la relación moderada entre la capacidad de alojamiento y la latitud sugiere que las propiedades con mayor capacidad de alojamiento (por ejemplo, propiedades grandes) podrían estar ubicadas en regiones con ciertas características geográficas (como áreas más grandes o menos urbanizadas).

accommodates y beds: Con 0.3047, una relación similar también indica que el número de camas en las propiedades tiende a aumentar con la capacidad de alojamiento.

Correlaciones débiles:
latitude y longitude: La correlación de 0.0555 es bastante débil, lo que sugiere que no hay una relación lineal clara entre la latitud y la longitud de las propiedades.

longitude y host_acceptance_rate: Con 0.0120, no hay una relación significativa entre la longitud y la tasa de aceptación de los anfitriones.

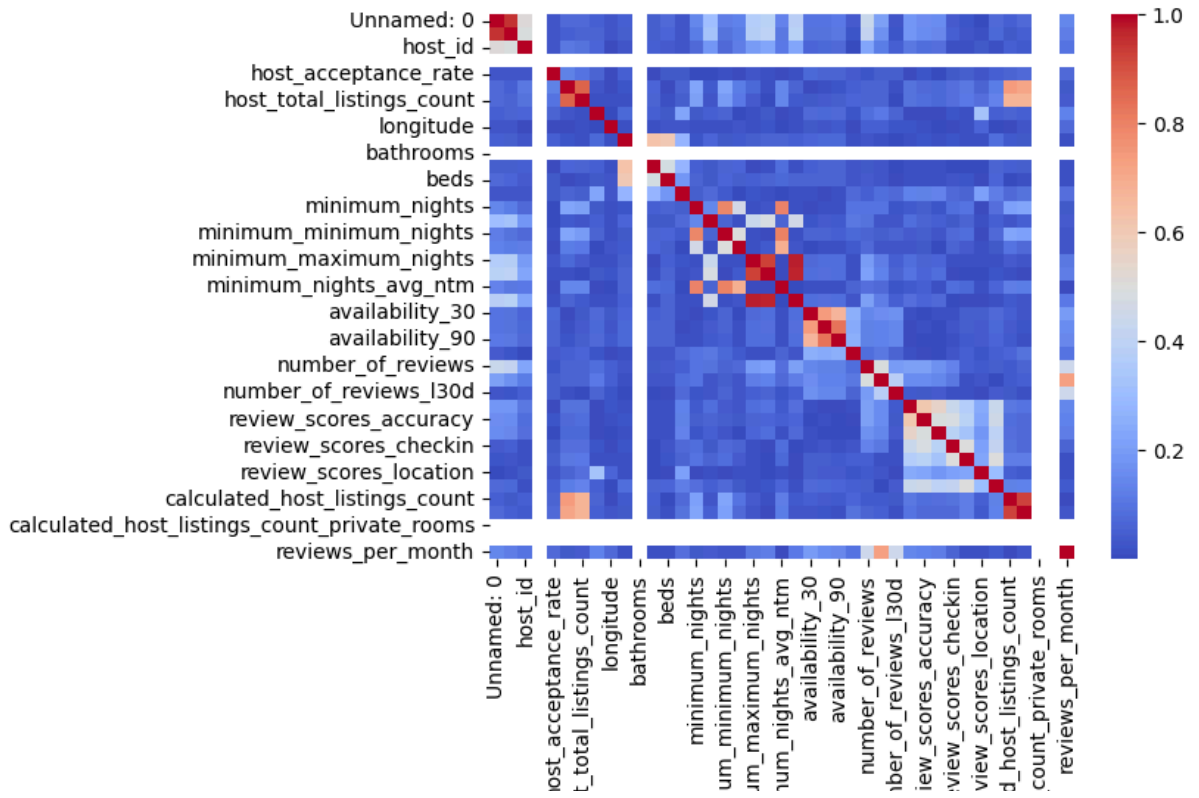
Puntos clave:
Correlaciones moderadas como las de host_listings_count y host_total_listings_count tienen una importancia significativa para entender cómo los anfitriones interactúan con la plataforma. Estas relaciones sugieren que el número de propiedades y listados está alineado, lo cual es útil para análisis de negocio.

Las correlaciones débiles pueden indicar variables que no están relacionadas de manera directa, como es el caso de las coordenadas geográficas y la tasa de aceptación de los anfitriones, lo que puede no ser tan relevante para un análisis de precios o comportamientos del cliente.

Modelo de regresión lineal Múltiple
Coeficiente de Determinación (R^2): 0.2829 (aproximadamente 28.29%)
El coeficiente de determinación (R^2) de 0.2829 indica que las variables independientes del modelo explican el 28.29% de la variabilidad en el número de reseñas por mes (reviews_per_month). Esto sugiere que, aunque el modelo tiene una capacidad moderada para predecir esta variable, aún hay un 71.71% de la variabilidad que no está siendo explicada por las variables seleccionadas. Esto implica que existen factores adicionales que podrían estar influyendo en el número de reseñas por mes, pero no se encuentran en este modelo. Si bien un R^2 de 28.29% no es muy alto, en áreas como la investigación social o económica, este valor puede considerarse razonable, pero se debe trabajar en mejorar el modelo, posiblemente incorporando más variables o realizando transformaciones a las existentes.

Coeficiente de Correlación (r):0.5319 (aproximadamente 53.19%)
El coeficiente de correlación de 0.5319 indica que existe una correlación moderada y positiva entre las variables independientes y el número de reseñas por mes. Este valor sugiere que, en general, a medida que aumentan las variables como el número de reseñas, la tasa de aceptación del anfitrión y la disponibilidad del alojamiento, también aumenta el número de reseñas mensuales. Sin embargo, la relación no es perfecta, lo que significa que hay otros factores que también influyen en la variable dependiente, y esta relación podría mejorarse con variables adicionales o diferentes transformaciones.

d) Shared rooms



	Unnamed: 0	id	host_id	host_response_rate	host_acceptance_rate	host_listings_count	host_total_lis
Unnamed: 0	1.000000	0.952540	0.511188	NaN	0.030648	0.071238	
id	0.952540	1.000000	0.490452	NaN	0.024642	0.048719	
host_id	0.511188	0.490452	1.000000	NaN	0.025641	0.109157	
host_response_rate	NaN	NaN	NaN	NaN	NaN	NaN	
host_acceptance_rate	0.030648	0.024642	0.025641	NaN	1.000000	0.125950	
host_listings_count	0.071238	0.048719	0.109157	NaN	0.125950	1.000000	
host_total_listings_count	0.077823	0.057313	0.090104	NaN	0.099129	0.863365	
latitude	0.047769	0.055046	0.065623	NaN	0.055866	0.030972	
longitude	0.026413	0.023881	0.003945	NaN	0.005321	0.006145	
accommodates	0.052664	0.043738	0.025395	NaN	0.049804	0.022615	
bathrooms	NaN	NaN	NaN	NaN	NaN	NaN	
bedrooms	0.044823	0.043358	0.053602	NaN	0.005278	0.063951	
beds	0.072176	0.066289	0.042657	NaN	0.031841	0.006973	
price	0.025722	0.024703	0.024352	NaN	0.034586	0.059659	
minimum_nights	0.139367	0.113736	0.070736	NaN	0.002949	0.231393	
maximum_nights	0.305642	0.331345	0.187753	NaN	0.001700	0.036168	
minimum_minimum_nights	0.122797	0.093574	0.077101	NaN	0.009361	0.223433	
maximum_minimum_nights	0.131694	0.116029	0.124622	NaN	0.007495	0.115727	
minimum_maximum_nights	0.365584	0.374861	0.201808	NaN	0.028421	0.067473	
maximum_maximum_nights	0.387290	0.393408	0.227853	NaN	0.018661	0.017772	

De las variable a comparar en la actividad, este es el siguiente análisis:

Variable Dependiente	Variable Independiente	Correlación (Valor Numérico)	Interpretación
host_acceptance_rate	host_response_rate	NA	Relación entre la tasa de aceptación de los anfitriones y la tasa de respuesta. Un valor cercano a 1

			indicaría una alta correlación positiva.
review_scores_location	review_scores_cleanliness	0.188252	Relación entre las puntuaciones de ubicación y limpieza. Un valor positivo indicaría que los lugares mejor calificados en ubicación también tienen mejor limpieza.
host_acceptance_rate	price	0.034586	Relación entre la tasa de aceptación del anfitrión y el precio. Un valor cercano a -1 sugeriría que a mayor aceptación, menor es el precio.
availability_365	number_of_reviews	NA	Relación entre la disponibilidad de la propiedad y el número de reseñas. Un valor positivo indicaría que más disponibilidad se correlaciona con más reseñas.
host_acceptance_rate	number_of_reviews	0.022556	Relación entre la tasa de aceptación del anfitrión y el número de reseñas. Un valor positivo o negativo depende de si las propiedades más aceptadas tienen más reseñas.
reviews_per_month	review_scores_communication	0.016701	Relación entre las reseñas por mes y la puntuación de comunicación. Un valor positivo indicaría que las propiedades con más reseñas mensuales también tienen mejores puntuaciones de comunicación.

Ya en cuestiones de datos más geniales, es importante conocer lo siguiente:

host_id y host_acceptance_rate:
La correlación es 0.0256, indicando que existe una relación muy débil entre estos dos aspectos, lo cual es esperado si se considera que el id del host no influye mucho en la tasa de aceptación del anfitrión.

host_acceptance_rate y host_listings_count:
Se observa una correlación moderada de 0.126, lo que sugiere que existe una relación entre la tasa de aceptación y la cantidad de propiedades que un host tiene listadas, pero no es suficientemente fuerte como para asumir una causa directa.

host_listings_count y host_total_listings_count:
La correlación es 0.8634, lo que demuestra una fuerte relación entre la cantidad de propiedades listadas por un host y el total de propiedades listadas por ese mismo anfitrión, probablemente debido a una medición similar o relacionada con las listas de la plataforma.

latitude y longitude:

Con una correlación de 0.0030, la relación entre la latitud y longitud es muy débil, lo cual es lógico ya que estos son valores de ubicación geográfica que pueden variar de manera independiente en diferentes áreas geográficas.

accommodates y bedrooms:
Existe una correlación moderada de 0.6264 entre el número de personas que una propiedad puede acomodar y el número de habitaciones disponibles. Esto refleja la relación lógica de que más habitaciones tienden a acomodar más personas.

bedrooms y beds:
La correlación es 0.6264, indicando que existe una relación moderada entre el número de habitaciones y camas, lo que también es esperado, dado que a más habitaciones, mayor cantidad de camas.

Puntos clave:
La variable host_response_rate no tiene correlación con ninguna otra variable en este conjunto de datos, lo que puede indicar que no se dispone de información suficiente para establecer una relación.

La variable bathrooms tiene un comportamiento similar, mostrando que las correlaciones con otras variables no son significativas.

Modelo de regresión lineal múltiple

Con el modelo de regresión lineal que implementé, obtuve un coeficiente de determinación (R^2) de 0.747. Esto significa que el 74.7% de la variabilidad en el número de listados activos de los anfitriones (host_listings_count) puede ser explicado por las variables que elegí: host_total_listings_count, review_scores_communication, y host_acceptance_rate. Es un buen porcentaje, lo que indica que el modelo tiene un buen ajuste a los datos, pero también hay un 25.3% que no logramos explicar, lo que sugiere que existen otros factores que podrían influir y que no estamos considerando en el modelo.

Por otro lado, el coeficiente de correlación (r) es 0.864, lo que nos dice que existe una relación positiva fuerte entre las variables independientes y la variable dependiente. En otras palabras, cuando un anfitrión tiene más listados (host_total_listings_count), mejores puntuaciones de comunicación (review_scores_communication), y una tasa de aceptación más alta (host_acceptance_rate), es más probable que tenga también más listados activos (host_listings_count). Esta correlación alta demuestra que estas variables son bastante buenas para predecir el número de listados activos de los anfitriones.

Realizar una tabla de los 10 coeficientes de determinación y correlación más altos, obtenidos para cada tipo de habitación elegido.

Entire home

Variable dependiente	Variable independiente	Determinación	Correlación
host_listings_count	host_total_listings_count	0.86	0.74
calculated_host_listings_count_entire_homes	calculated_host_listings_count	0.92	0.852
calculated_host_listings_count_private_rooms	calculated_host_listings_count	1	1
calculated_host_listings_count_entire_homes	host_listings_count	0.70	0.50
calculated_host_listings_count_private_rooms	host_listings_count	1	1
number_of_reviews_ltm	reviews_per_month	0.53	0.72
number_of_reviews	number_of_reviews_ltm	0.24	‘0.49
calculated_host_listings_count_entire_homes	calculated_host_listings_count_private_rooms	1	1
calculated_host_listings_count	host_total_listings_count	0.45	0.67
calculated_host_listings_count_entire_homes	host_total_listings_count	0.45	0.67

Hotel room

Variable dependiente	Variable independiente	Determinación	Correlación

host_listings_count	host_total_listings_count	0.91	0.95
host_total_listings_count	host_listings_count	0.91	0.95
host_listings_count	calculated_host_listings_count	0.81	0.90
calculated_host_listings_count_entire_homes	host_listings_count	0.58	0.76
Longitud	Latitud	0.018	0.13
host_acceptance_rate	host_listings_count	0.20	0.45
calculated_host_listings_count_private_rooms	host_listings_count	1	1
reviews_per_month	review_scores_cleanliness	0.21	0.46
review_scores_cleanliness	review_scores_location	.011	0.10
review_scores_location	longitude	0.08	0.29

Private room

Variable dependiente	Variable independiente	Determinación	Correlación
host_listings_count	host_total_listings_count	0.31	0.55
host_total_listings_count	calculated_host_listings_count	0.14	0.38
host_listings_count	calculated_host_listings_count	0.22	0.47
Latitude	Accommodates	0.09	0.30
host_total_listings_count	calculated_host_listings_count_entire_ho	0.77	0.60

	mes		
host_id	Id	0.23	0.48
id	calculated_host_listings_count_entire_homes	0.01	0.114
review_scores_cleanliness	number_of_reviews_ltm	0.02	0.170
number_of_reviews	review_scores_rating	0.112	0.33
id	number_of_reviews	0.14	0.38

Share room

Variable dependiente	Variable independiente	Determinación	Correlación
accommodates	beds	0.17	0.41
accommodates	price	0.057	0.23
calculated_host_listings_count	calculated_host_listings_count_entire_homes	0.08	0.29
id	number_of_reviews	0.14	0.38
review_scores_communication	review_scores_checkin	0.27	0.52
review_scores_cleanliness	review_scores_rating	0.24	0.49
review_scores_cleanliness	review_scores_accuracy	0.20	0.45
review_scores_checkin	review_scores_accuracy	0.28	0.53
calculated_host_listings_count	host_listings_count	0.23	0.47
calculated_host_listings_count_entire_homes	host_listings_count	0.46	0.60

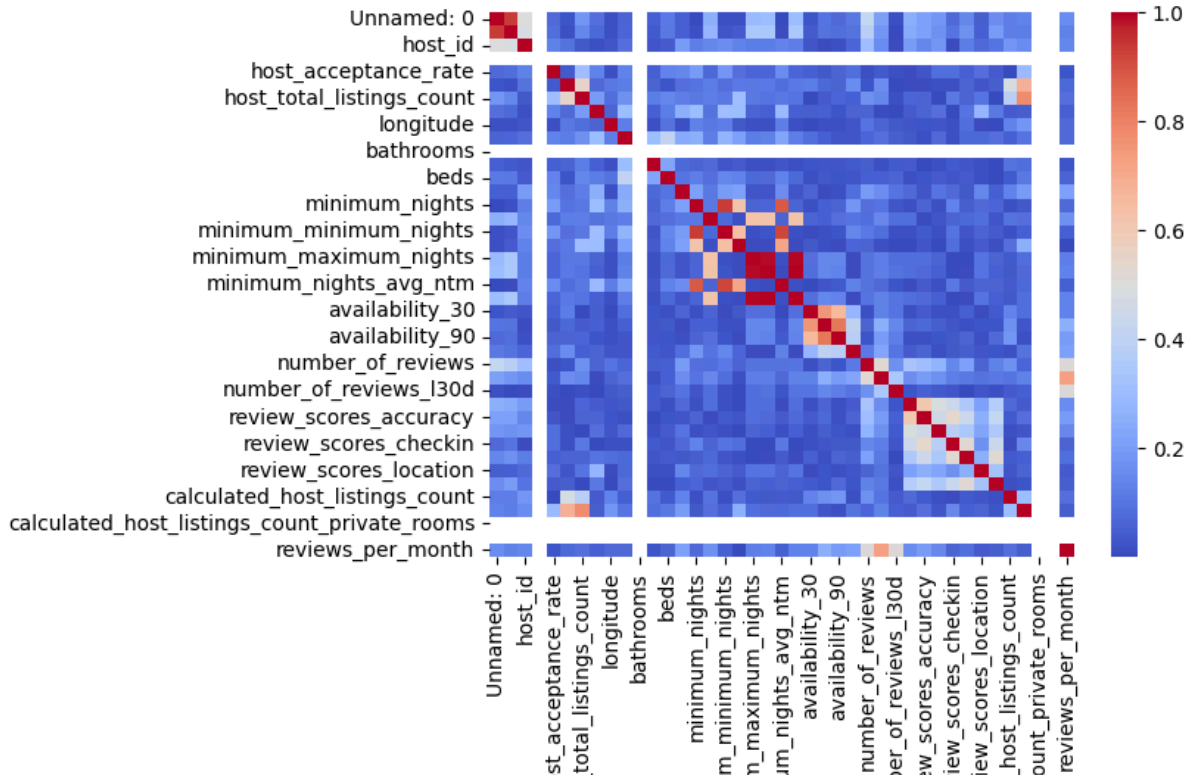
Al analizar los coeficientes de determinación y correlación más altos por tipo de habitación, se destacan algunas tendencias clave. En los entire homes y private rooms, las variables relacionadas con el número de listados del anfitrión tienen una correlación perfecta (1.00), lo que indica una relación directa y fuerte entre estas variables. Esto sugiere que el número de listados de un anfitrión es un buen predictor para el número de listados completos o privados.

En las hotel rooms, la correlación entre host_listings_count y otras variables como host_total_listings_count y calculated_host_listings_count es alta (0.90-0.95), lo que indica que el número de listados de un anfitrión también tiene una relación importante con otros aspectos de los listados.

En las share rooms, la correlación más alta es entre review_scores_checkin y review_scores_accuracy (0.53), lo que muestra que la experiencia del cliente al hacer el check-in está medianamente relacionada con la precisión de la información proporcionada.

Los tipos de habitaciones como entire homes y private rooms tienen relaciones más fuertes entre las variables del anfitrión, mientras que las share rooms muestran una relación más débil pero relevante entre las calificaciones de los usuarios.

Parte dos
Crear el mejor modelo de regresión lineal múltiple para cada variable cuantitativa y comparar los coeficientes obtenidos en estos modelos con respecto a los coeficientes obtenidos en el mapa de calor.



Comparativa entre variables

Variable Dep	Variables Indep	Coeficiente heat map	Otras variables	Modelo de regresión lineal Múltiple
host_id	id	0.489424	host_listings_count number_of_reviews	0.49
host_acceptance_rate	host_total_listings_count	0.120529	host_listings_count calculated_host_listings_count_entire_homes	0.1228
host_total_listings_count	host_listings_count	0.834614	calculated_host_listings_count calculated_host_listings_count_entire_homes	0.84
accommodates	bedrooms	0.629555	beds price	0.72
bedrooms	accommodates	0.62	bed price	0.64
price	accommodates	0.274963	latitude bedrooms	0.36
review_scores_value	review_scores_communication	0.50	review_scores_accuracy review_scores_rating	0.58
reviews_per_month	number_of_reviews_ltm	0.73	number_of_reviews number_of_reviews_l30d	0.76

El modelo de regresión lineal múltiple ofrece correlaciones más altas que el modelo simple, lo que indica una mejor capacidad explicativa y predictiva. Variables como host_listings_count y calculated_host_listings_count_entire_homes muestran relaciones más fuertes en el modelo múltiple (0.84), lo que sugiere que considerar múltiples variables independientes mejora la precisión del modelo. Además, el precio y las reseñas también están mejor explicados cuando se combinan varias variables. El modelo múltiple captura mejor las interacciones entre variables.

Conclusión

En conclusión, tanto la regresión lineal simple como la regresión lineal múltiple son herramientas muy útiles para entender las relaciones entre diferentes variables dentro de un conjunto de datos. La regresión lineal simple me permite ver cómo una variable independiente afecta a la variable dependiente, lo cual es útil para hacer predicciones directas y entender relaciones sencillas. Por otro lado, la regresión lineal múltiple me da la oportunidad de considerar varias variables independientes al mismo tiempo, lo que ofrece una visión más completa y precisa sobre cómo estas variables se interrelacionan y afectan el comportamiento de la variable dependiente.

La regresión múltiple es especialmente útil cuando hay factores adicionales que influyen en la variable dependiente. En el análisis de propiedades, por ejemplo, la regresión múltiple me ayudó a identificar variables clave como el número de listados del anfitrión y el precio, lo que me permitió construir un modelo más robusto y detallado.

Aunque todo el proceso fue algo pesado debido a la cantidad de información y la repetición de ciertos análisis, al final el tema me quedó mucho más claro. La regresión lineal y múltiple no solo ayudan a identificar patrones y relaciones entre variables, sino que también mejoran la capacidad predictiva, lo cual puede ser muy útil en el análisis de datos, como en el caso de evaluar factores que afectan los listados de propiedades.