Actividad 1.

Regresión Lineal Simple y Múltiple.

Gestión de proyectos de plataformas tecnológicas
Gpo 201

04 de Octubre, 2025
ITESM Puebla

Valeria Becerril Hernandez | A01736860

Introducción.

El presente análisis explora la dinámica del mercado de alojamientos a partir de una base de datos robusta, que requirió un riguroso proceso de limpieza y transformación. La fase inicial incluyó la selección de 47 columnas clave, la conversión de tipos de datos a formato numérico (incluyendo la binarización de host_is_superhost) y el tratamiento de valores atípicos mediante el método de Rango Intercuartílico (IQR), técnica elegida por su solidez. El núcleo del estudio se centró en dos fases analíticas: la Regresión Simple, para comprender las relaciones lineales entre variables segmentadas por tipo de alojamiento (room_type), y la Regresión Múltiple, para construir modelos predictivos para métricas clave como el precio y las puntuaciones de reseña. Los resultados revelan una multicolinealidad estructural significativa en las variables de gestión y tamaño, lo que subraya la necesidad de un análisis segmentado para capturar las dinámicas de mercado únicas de cada tipo de alojamiento.

Limpieza de datos.

Tomando la base de datos de México y explorando, nos arroja que cuenta con 26,401 filas por 79 columnas, lo cual nos indica que estamos que es una base de datos robusta. Por otro lado se encontró que la base de datos cuenta con 206,551 valores nulos en 42 filas del data frame. Para seguir con el proceso de limpieza de valores nulos y valores atípicos y para hacer de manera más eficiente, se creó una copia de la base de datos con solo las columnas de interés para esta práctica, es decir, las columnas siguientes:

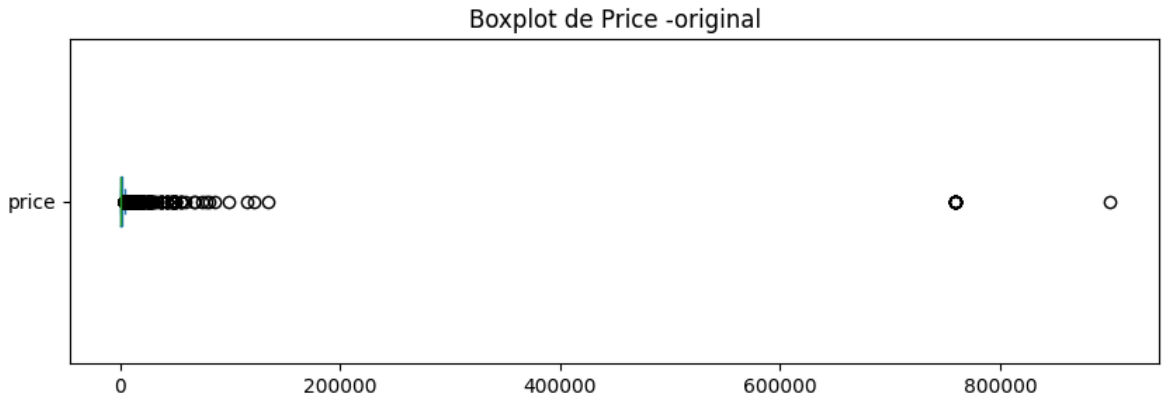
Columna	Detalle
room_type	Tipo de habitación, se clasifica en 4 tipos: Entire home/apt, Private room, Hotel room y Shared room.
host_is_superhost	Si se considera como un super anfitrión (evaluado en ‘t’ o ‘f’)
host_acceptance_rate	Tasa en la que el anfitrión acepta solicitudes de reserva (se evalúa en porcentaje)
host_response_rate	Tasa de respuesta del anfitrión (se evalúa en porcentaje)
price	Precio del hospedaje en la moneda local
number_of_reviews	Número de reseñas
review_scores_rating	Puntuación general del anuncio (escala de evaluación de 1-5)
calculated_host_listings_count	Número de anuncio que tienen el anfitrión en la ciudad o región geográfica

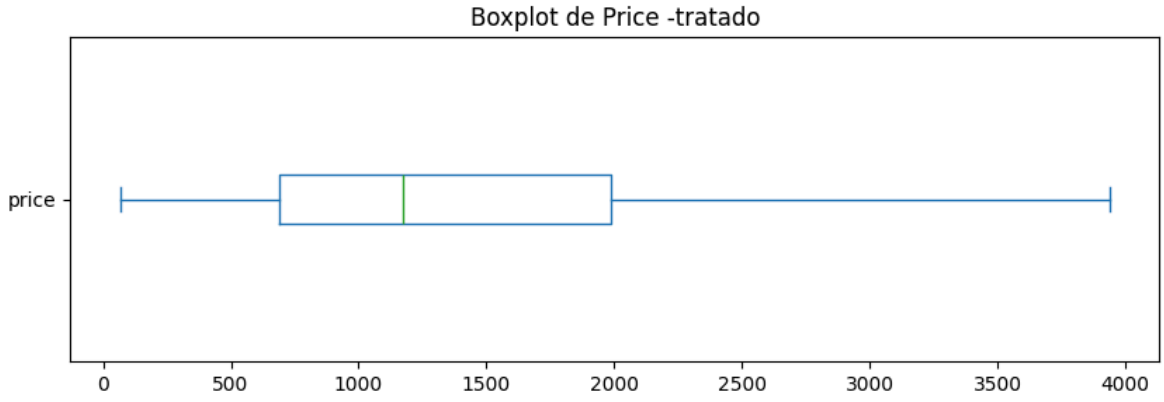
availability_365	Disponibilidad dentro de los 365 días
reviews_per_month	Número promedio de reseñas por mes que tienen el anuncio a lo largo de su vida útil
review_scores_communication	Puntuación de la comunicación con el anfitrión (escala de evaluación de 1-5)

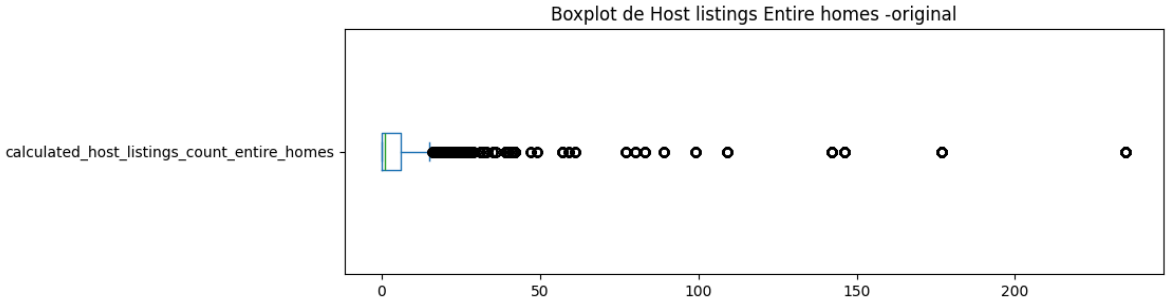
Gracias a que vamos a realizar un análisis de modelos, se optó por crear una base de datos con variables cuantitativas y dejando una única variable cualitativa (‘room_type’). Por lo cual, la nueva base de datos se conforma por 26,401 filas por 50 columna, sin embargo sufrió un cierto ajuste, para iniciar las columnas ‘price’, ‘host_response_rate’, ‘host_acceptance_rate’, ‘host_is_superhost’ se ajustaron de tipo objeto a categoría. A ‘price’ se le quito el signo de pesos, las comas y los puntos, para ‘host_response_rate’ y ‘host_acceptance_rate’ se quito el signo de porcentaje y se divido sobre 100 para que un valor numérico sin porcentaje y por último para la columna ‘host_is_superhost’ se sustituyó la ‘t’ (true/ verdadero) a 1 y ‘f’ (false/falso) a 0. Otro cambio que se realizó, fue la eliminación de 3 columnas, ‘neighbourhood_group_cleansed’, ‘calendar_updated’ y ‘license’, ya que contaban con toda la columna en sin información. Quedando al final con 26,401 filas por 47 columnas.

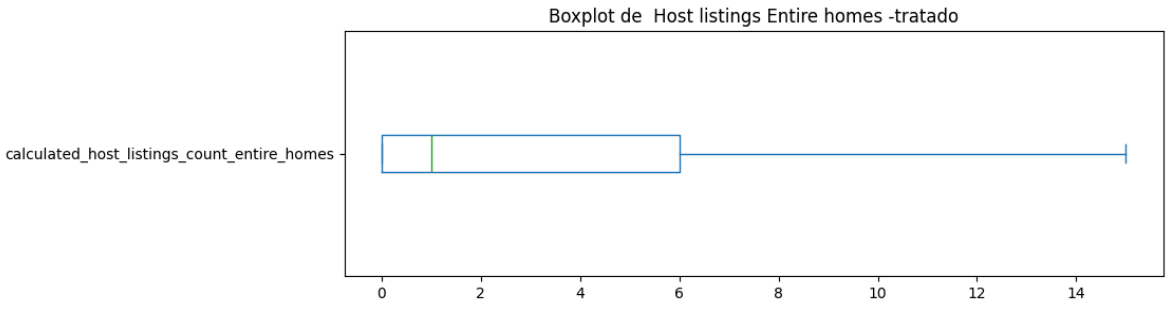
Para tratar los valores nulos y los outliers se utilizó la combinación de la media para la imputación de valores nulos y el Rango Intercuartílico (IQR) para el tratamiento de outliers, ya que se considera el mejor método en el análisis de datos exploratorio, porque logran un equilibrio óptimo entre simplicidad y robustez. La media es la técnica de imputación más rápida y eficiente, garantizando que el promedio general de la distribución se preserve y que no haya pérdida de datos, lo cual es esencial para que los cálculos estadísticos posteriores se ejecuten sin errores. Por su parte, el método IQR es el más robusto para identificar outliers porque se basa en los cuartiles (Q1 y Q3), los cuales no se ven afectados por los valores extremos, a diferencia de la desviación estándar y la media; al combinarlo con la winsorización (.clip()), se consigue un tratamiento efectivo que ajusta los valores extremos sin eliminarlos, manteniendo así la integridad del conjunto de datos.

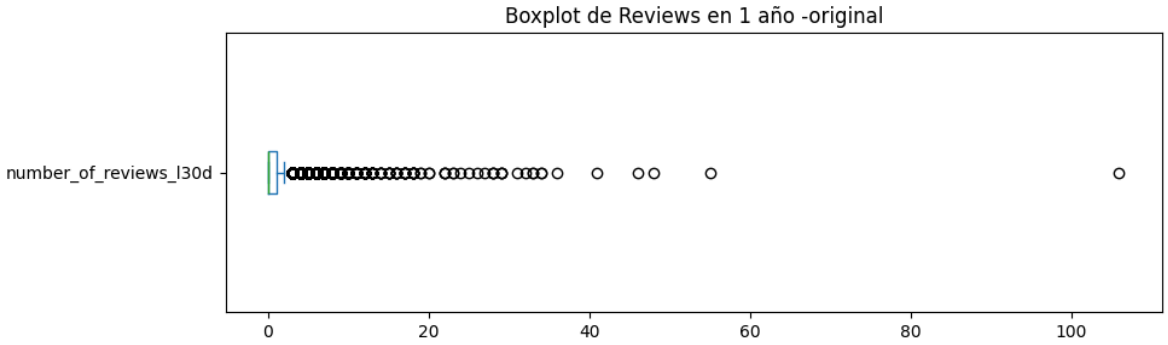
El top 5 de las variables que más sufrieron el ajuste de los outliers son: ‘price’, ‘calculated_host_listings_count_entire_homes’, ‘number_of_reviews_l30d’, ‘minimum_nights’ y ‘calculated_host_listings_count_private_rooms’

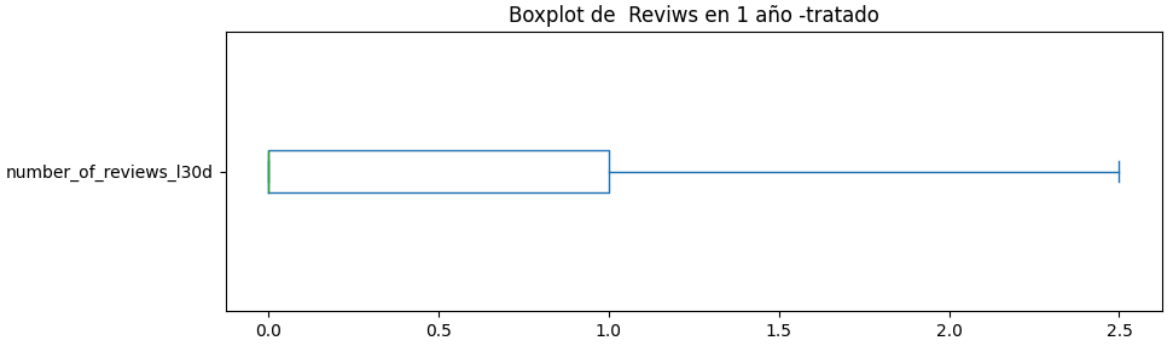

Boxplot de Price -original
price
0
200000
400000
600000
800000

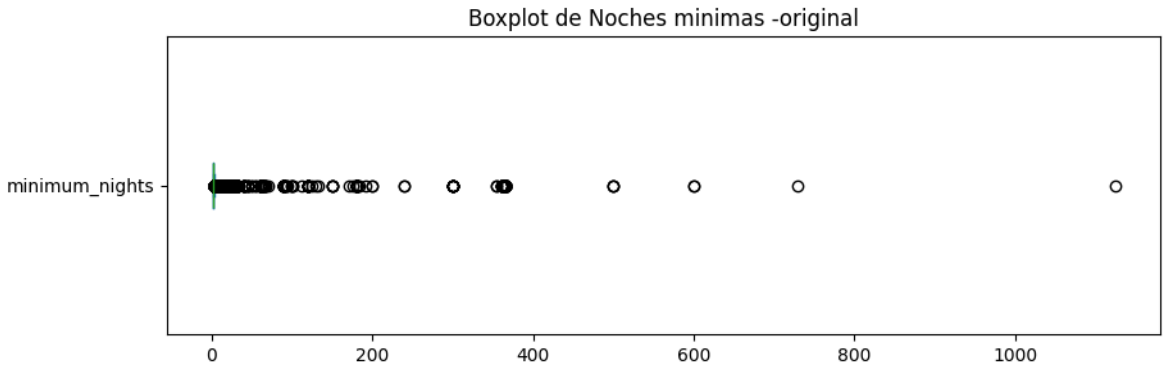

Boxplot de Price -tratado
price
0
500
1000
1500
2000
2500
3000
3500
4000

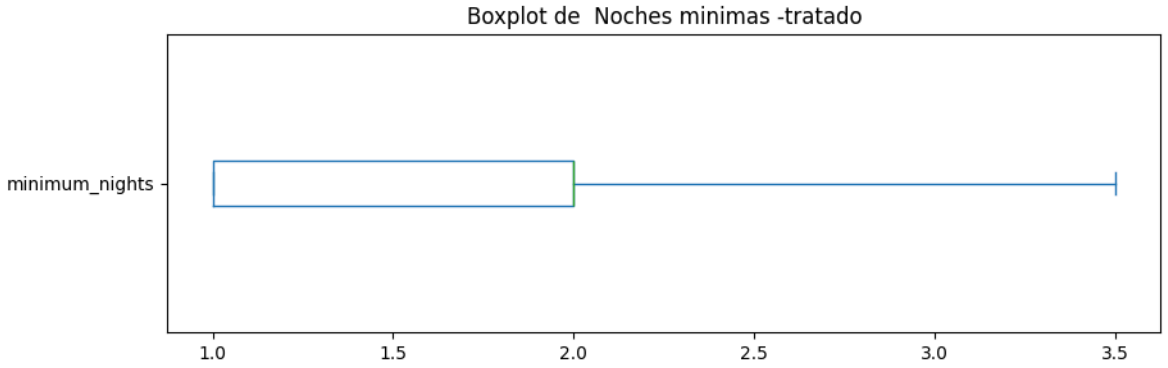

Boxplot de Host listings Entire homes -original
calculated_host_listings_count_entire_homes
0
50
100
150
200

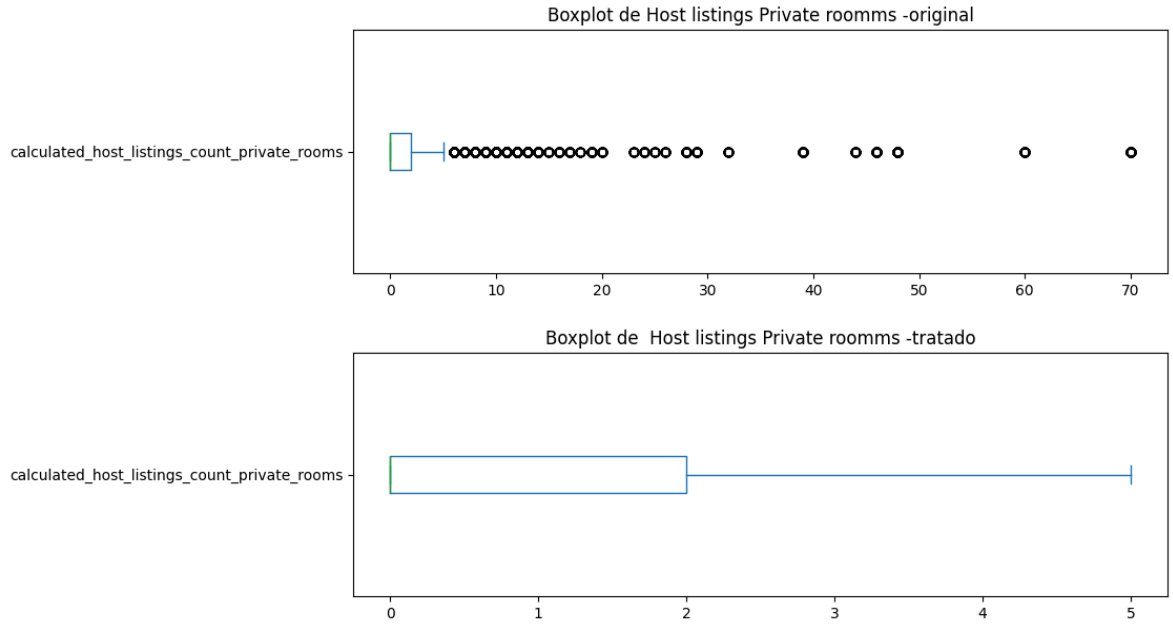

Boxplot de Host listings Entire homes -tratado
calculated_host_listings_count_entire_homes
0
2
4
6
8
10
12
14


Boxplot de Reviews en 1 año -original
number_of_reviews_l30d
0
20
40
60
80
100


Boxplot de Reviws en 1 año -tratado
number_of_reviews_l30d
0.0
0.5
1.0
1.5
2.0
2.5


Boxplot de Noches minimas -original
minimum_nights
0
200
400
600
800
1000


Boxplot de Noches minimas -tratado
minimum_nights
1.0
1.5
2.0
2.5
3.0
3.5

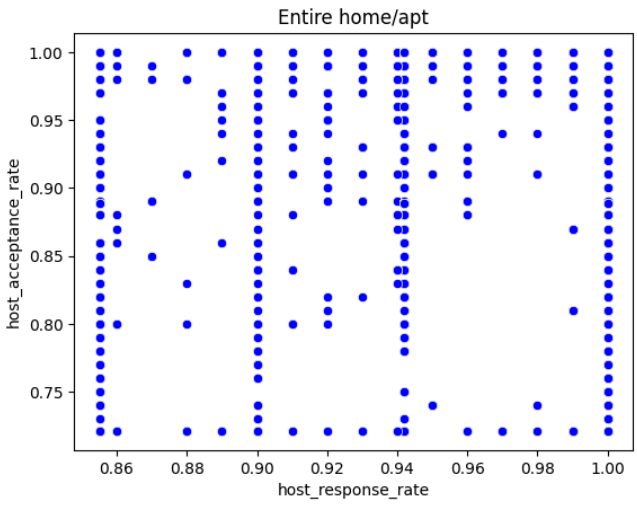


Regresión Simple.

Para iniciar con el análisis de la correlación que existe en cada tipo de habitación con respecto a las variables indicadas más adelantes, conocemos cómo se componen los cuatro tipos de habitaciones.

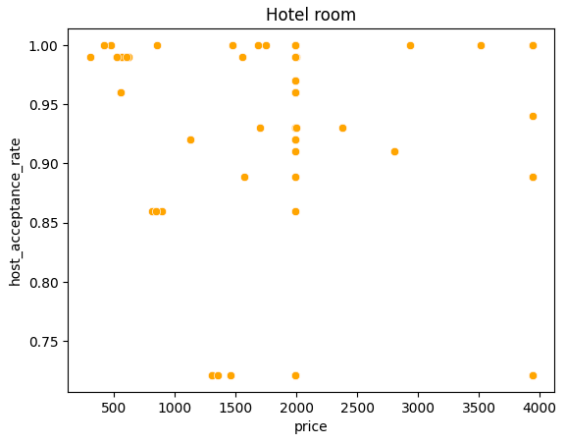
Tipo de habitación	Número de variables
Entire home/apt	17,235
Private room	8,867
Shared room	208
Hotel room	91

Luego de la observación de cómo se compone se dividieron en 4 bases de datos que se filtraron para hacer una comparación de cómo se tiene la correlación.

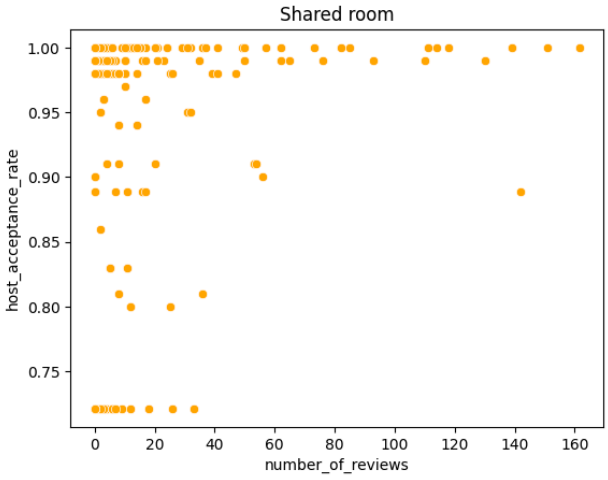


Primer bloque 'host_acceptance_rate' (dependiente) vs 'host_response_rate' (independiente), el tipo de habitación 'Entire home/apt' cuenta con 0.510 de correlación mientras que el menor fue 'Hotel room' con -0.066, esto nos indica que la relación lineal entre la tasa de respuesta y la tasa de aceptación del anfitrión es

significativamente diferente según el tipo de alojamiento. En el caso de 'Entire home/apt', el coeficiente muestra una correlación positiva moderada a fuerte, lo que sugiere que a medida que la tasa de respuesta del anfitrión aumenta (responde más rápido o a más solicitudes), también tiende a aumentar su tasa de aceptación de reservas. Por otro lado, el coeficiente de 'Hotel room' es prácticamente nulo, indicando que no existe una relación lineal entre la velocidad o frecuencia de respuesta del hotel y si acepta o rechaza una reserva. Esta clara diferencia subraya la importancia de segmentar el análisis, ya que las dinámicas de negocio y las variables que influyen en la aceptación son completamente distintas entre un anfitrión individual y la gestión estandarizada de un hotel.

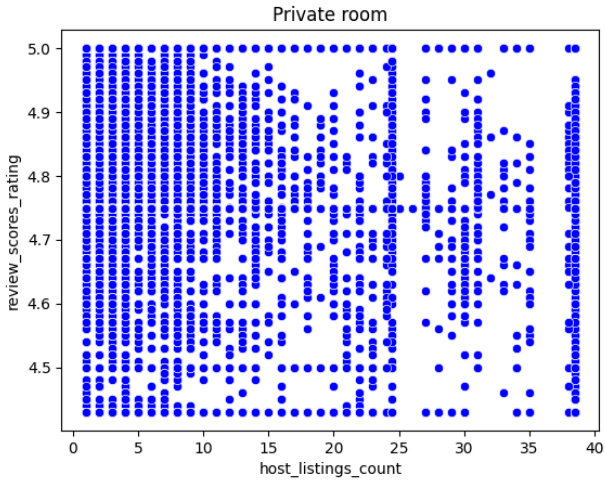


Segundo bloque 'host_acceptance_rate' (dependiente) vs 'price' (independiente), teniendo en cuenta que la correlaciones estuvieron muy bajas, siendo:
Entire home/apt -0.003, Private room 0.046, Shared room -0.095 y Hotel room -0201. Esto nos indica que el precio del alojamiento tiene una relación lineal negligible o muy débil con la tasa de aceptación de reservas del anfitrión en la mayoría de los casos, con la excepción notable de las habitaciones de hotel.
Específicamente, para 'Entire home/apt' (−0.003) y 'Private room' (0.046), la correlación es prácticamente cero. Esto sugiere que, para los anfitriones individuales, la decisión de aceptar o rechazar una reserva no está influenciada significativamente por el precio de su listado. Es decir, un precio más alto o más bajo no implica que el anfitrión sea más o menos propenso a aceptar la solicitud. La mayor correlación negativa se observa en 'Hotel room' (−0.201), lo que, aunque sigue siendo débil, sugiere una tendencia: los precios más altos en las habitaciones de hotel se asocian con una menor tasa de aceptación (o una mayor tasa de rechazo/cancelación). Esta dinámica es inversamente opuesta a la intuición de un mercado individual y podría reflejar políticas de gestión de inventario más rígidas por parte de los hoteles, donde las solicitudes de precio más alto (quizás para temporadas pico) son rechazadas por falta de disponibilidad o por sistemas de gestión externos. Esta segmentación es fundamental, ya que confirma que las estrategias de aceptación o rechazo están impulsadas por factores internos del negocio, no por el precio en sí mismo.

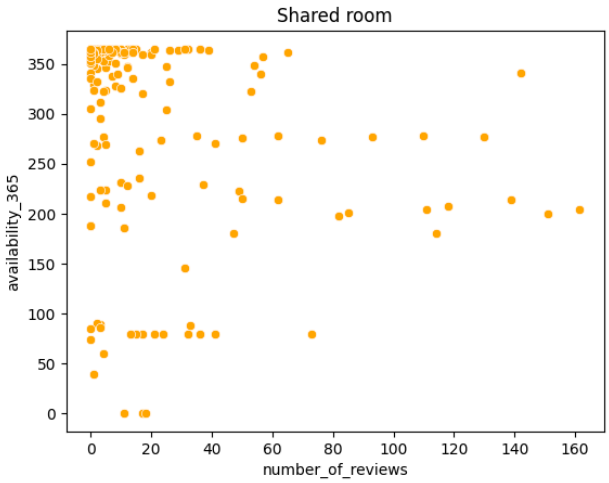


Tercer bloque 'host_acceptance_rate' (dependiente) vs 'number_of_reviews' (independiente). Los tipos 'Private room' (0.288), 'Hotel room' (0.284), y 'Entire home/apt' (0.273) muestran coeficientes muy similares. Esta uniformidad sugiere que, en general, una mayor cantidad de reseñas (indicando experiencia, historial y confianza en la plataforma) está débilmente asociada con una mayor

propensión del anfitrión a aceptar las solicitudes de reserva. En esencia, una trayectoria probada en la plataforma actúa como un factor de confianza. La única variación significativa es 'Shared room' (0.146), cuya correlación es notablemente menor. Esto implica que, para los alojamientos compartidos, la cantidad de reseñas históricas del listado tiene mucha menos influencia en la tasa de aceptación del anfitrión. Esta segmentación confirma que, si bien la experiencia es un factor de aceptación en todo el mercado, su impacto es menor en el nicho de las habitaciones compartidas.

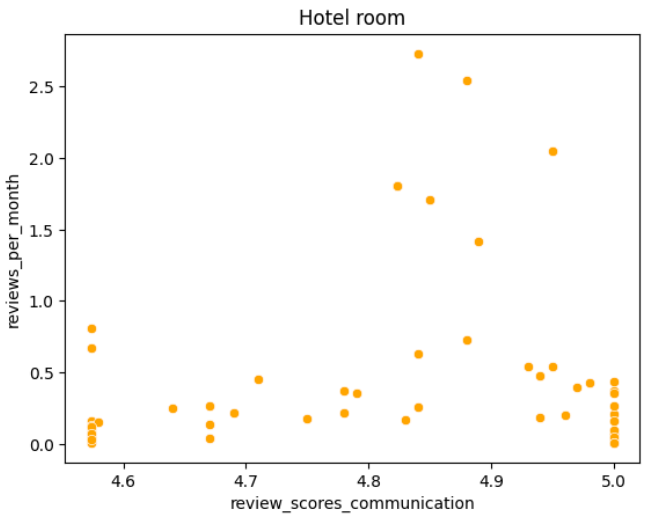


Cuarto bloque 'review_scores_rating' (dependiente) vs 'host_listings_count' (independiente). Esto nos indica que existe una correlación negativa consistente, aunque débil, en todos los tipos de alojamiento entre la cantidad total de listados que tiene un anfitrión y la puntuación general de las reseñas que recibe ese listado. Las correlaciones son muy similares en magnitud (alrededor de −0.12 a −0.17), lo que sugiere que la dinámica es común en todo el mercado, independientemente de si se trata de una casa completa o una habitación privada. La relación negativa implica que, a medida que la cantidad de propiedades gestionadas por un anfitrión aumenta, la puntuación general de las reseñas tiende a disminuir ligeramente. Este hallazgo es crucial para entender la gestión de la calidad: un anfitrión que opera una sola propiedad o pocas (con un host_listings_count bajo) tiende a ofrecer un nivel de atención y calidad percibida que resulta en puntuaciones marginalmente mejores. En contraste, los anfitriones o empresas que gestionan un gran volumen de propiedades a menudo se enfrentan a desafíos de estandarización y atención personalizada, lo que se refleja en una leve pero observable penalización en las puntuaciones de calidad otorgadas por los huéspedes. Esta segmentación confirma que la gestión a gran escala tiene un costo pequeño pero medible en la satisfacción general del huésped.



Quinto bloque 'availability_365' (dependiente) vs 'number_of_reviews' (independiente). Esto nos indica que, para la mayoría de las categorías, la relación lineal entre cuántos días está disponible un listado anualmente y la cantidad de reseñas que tiene es casi nula. Los coeficientes de 'Entire home/apt' (0.030) y 'Private room' (0.079) están muy cerca de cero, sugiriendo que tener un listado disponible por más días no se traduce directamente en un mayor o menor número de reseñas.

La excepción más notable es 'Shared room' (−0.277), que muestra la correlación negativa más fuerte de este bloque. Esto implica que, para las habitaciones compartidas, los listados que están disponibles por más días al año tienden a tener un menor número de reseñas. Esto puede ser un indicio de que los listados de habitaciones compartidas con alta disponibilidad son menos demandados o que son propiedades nuevas, que han estado activas poco tiempo o han tenido bajas tasas de ocupación a pesar de estar abiertas. Por último, 'Hotel room' (0.137) tiene una correlación positiva débil, lo que podría sugerir que la disponibilidad constante contribuye ligeramente a acumular más reseñas. La segmentación es esencial aquí, ya que revela que las dinámicas operativas de las habitaciones compartidas son inversas a las del resto del mercado.

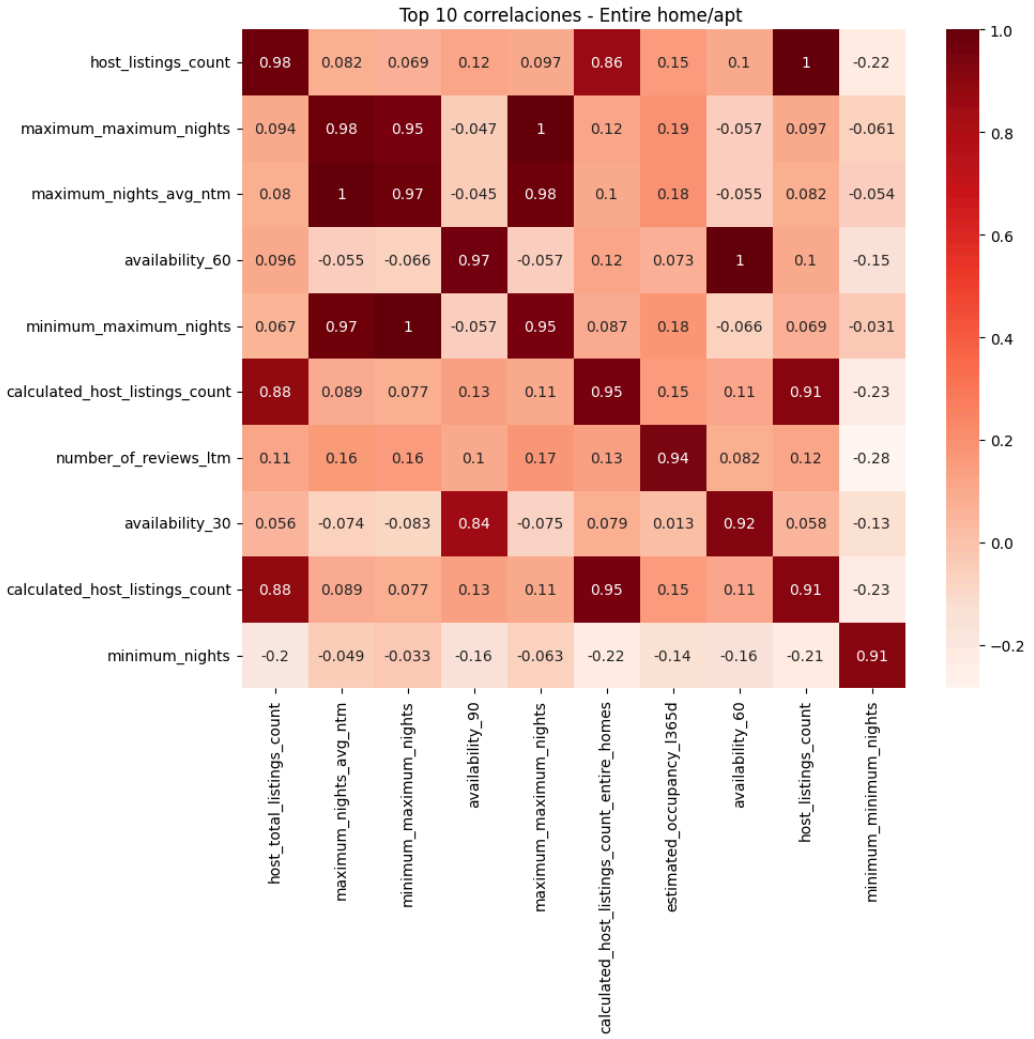


Sexto bloque 'reviews_per_month' (dependiente) vs 'review_scores_communication' (independiente). Esto nos indica que, en todas las categorías, la relación lineal entre la calidad de la comunicación del anfitrión y la frecuencia con la que un listado recibe reseñas es prácticamente nula. Todos los coeficientes están muy cerca de cero (entre −0.05 y 0.04), lo que se interpreta como una ausencia de correlación. Este hallazgo es crucial porque refuta una hipótesis común: que una excelente comunicación del anfitrión se traduciría directamente en una mayor tasa de reseñas (mayor reviews_per_month). El análisis segmentado demuestra que la puntuación de comunicación, aunque es importante para la calidad general del servicio, no es el factor principal que impulsa a los huéspedes a escribir reseñas. En esencia, la frecuencia de las reseñas parece estar impulsada por factores más relacionados con la tasa de ocupación o las políticas específicas del anfitrión para solicitar feedback, y no por la percepción de si la comunicación fue buena o mala. La segmentación confirma que esta falta de relación es una dinámica uniforme en todo el mercado, aplicándose tanto a anfitriones individuales como a alojamientos gestionados por hoteles.

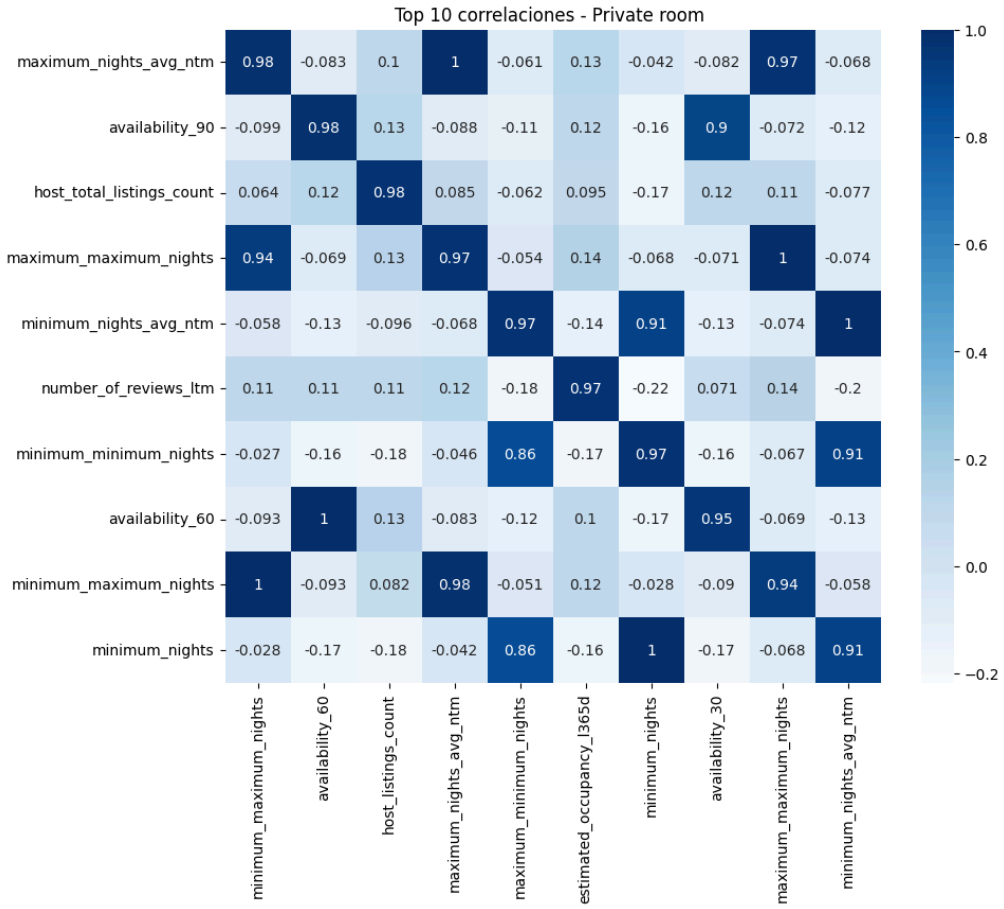
Tabla de las 10 variables con mayor correlación.

Para el tipo de cuarto **‘Entire home/apt’** el proceso de análisis revela una alta multicolinealidad entre las variables predictoras para los alojamientos, lo cual es la principal conclusión de este Top de correlaciones. Las relaciones más fuertes, muchas de ellas superando el 0.90 (como entre host_total_listings_count y host_listings_count, o entre las métricas de noches máximas), indican que estas variables miden conceptos idénticos o casi idénticos dentro de la gestión del inventario y las políticas de reserva; esta redundancia extrema significa que para construir un modelo predictivo efectivo y estable, es necesario eliminar las variables duplicadas (p. ej., mantener solo una de las métricas de conteo de

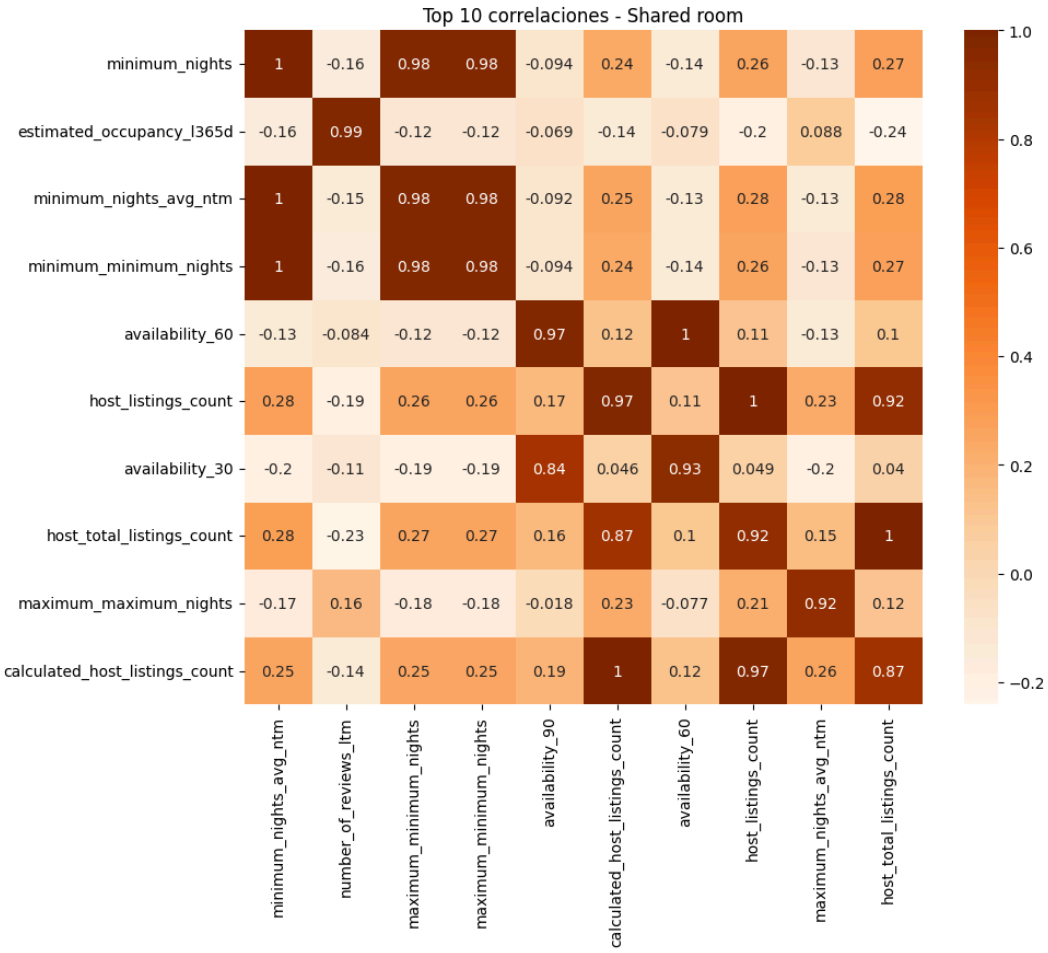
listados) para evitar que la alta interdependencia distorsione y complique la interpretación de los coeficientes de regresión



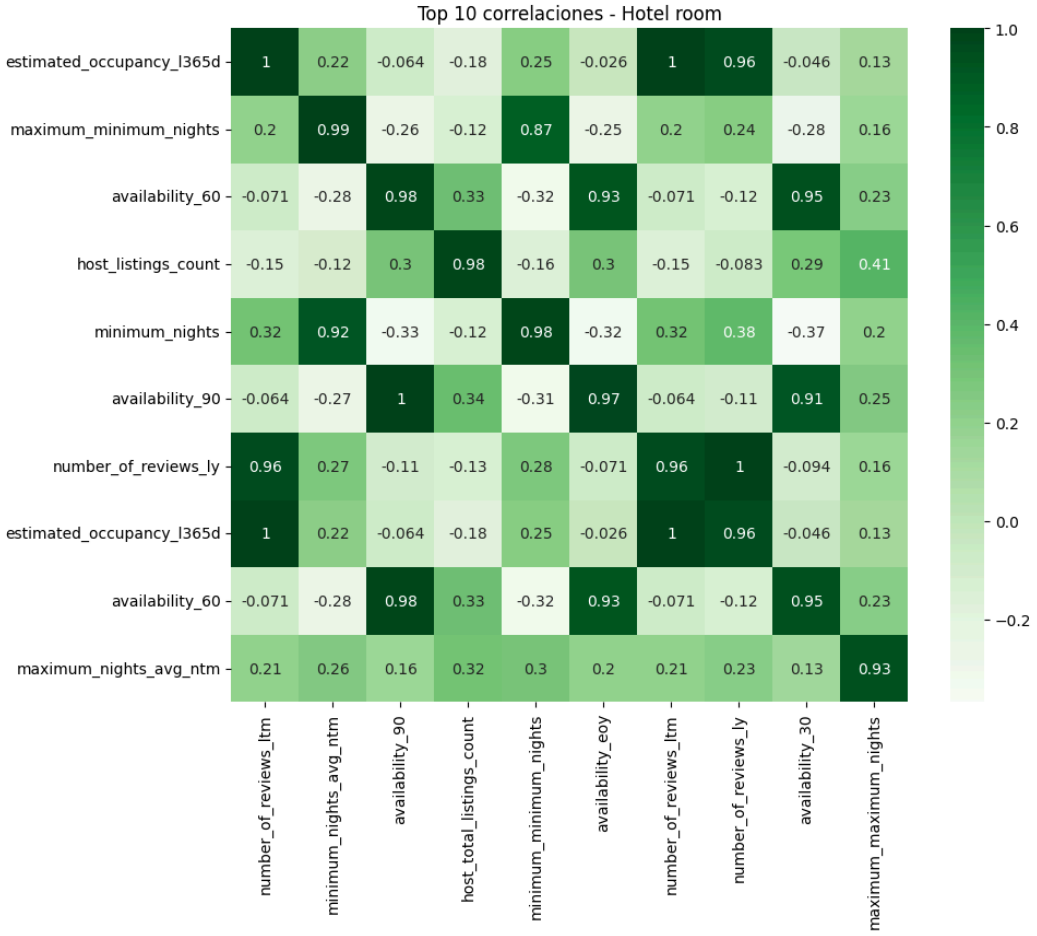
El proceso de análisis para los alojamientos de tipo **'Private room'** revela patrones de correlación absoluta muy similares a los observados en 'Entire home/apt', lo que subraya una alta redundancia métrica y la presencia de multicolinealidad entre las variables. Las correlaciones más fuertes (muchas por encima de 0.95) se centran en tres ejes principales: la gestión de las reglas de noches de estancia, las métricas de disponibilidad de calendario y los conteos de inventario del anfitrión. Por ejemplo, la relación de 0.982 entre maximum_nights_avg_ntm y minimum_maximum_nights indica que las políticas de estancia máxima son prácticamente idénticas entre los listados de un mismo anfitrión, y el 0.979 entre availability_90 y availability_60 confirma la consistencia en la gestión del calendario a corto y medio plazo. Asimismo, el valor de 0.967 entre estimated_occupancy_l365d y number_of_reviews_ltm muestra una fuerte dependencia entre la ocupación estimada y las reseñas recientes. En conclusión, esta segmentación confirma que para el modelado predictivo, es crucial reducir el número de variables de cada uno de estos ejes (reglas de estancia, disponibilidad, e inventario) para eliminar la redundancia y asegurar la estabilidad de los coeficientes de regresión.



El análisis para los alojamientos de tipo **'Shared room'** (Habitación Compartida) revela un patrón de redundancia métrica aún más extremo que en las categorías anteriores, con las primeras siete correlaciones alcanzando valores excepcionalmente altos, casi perfectos, que superan el 0.97. Esto subraya una severa multicolinealidad dentro de este nicho de mercado. La máxima correlación es 1.00 entre minimum_minimum_nights y minimum_nights, lo que significa que en la práctica, estas dos variables son idénticas en todos los listados de habitaciones compartidas. Este patrón se repite en las reglas de estancia mínima (minimum_nights_avg_ntm con sus contrapartes), demostrando que las políticas de estancia mínima son fijas y uniformes para los anfitriones de habitaciones compartidas. Además, el fuerte lazo entre las métricas de inventario (por ejemplo, 0.971 entre calculated_host_listings_count y host_listings_count) y la alta correlación en disponibilidad (0.972 entre availability_90 y availability_60) confirman que, al igual que en otros tipos de alojamiento, hay duplicidad en la medición de la gestión y el calendario. La alta correlación entre estimated_occupancy_l365d y number_of_reviews_ltm (0.985) es también notable, sugiriendo que la ocupación estimada se traduce casi directamente en la acumulación de reseñas recientes. En consecuencia, cualquier modelo predictivo para 'Shared room' requeriría una agresiva eliminación de variables redundantes para aislar los factores causales verdaderos y mantener la estabilidad del modelo.



El análisis de correlaciones para los alojamientos tipo **'Hotel room'** sigue el patrón de extrema redundancia métrica observado en las otras categorías, pero presenta una característica única que subraya su gestión profesional: la identidad casi perfecta entre las métricas de demanda y feedback. La principal conclusión es la multicolinealidad severa en las variables de gestión. Las primeras correlaciones, incluyendo el 1.00 entre estimated_occupancy_l365d y number_of_reviews_ltm, demuestran que, en la práctica, la ocupación estimada es una medida directa de las reseñas recientes, lo que sugiere un sistema de gestión altamente eficiente donde cada estadía se traduce en un feedback medible. Este patrón de duplicidad métrica se extiende a otros ejes de gestión: el 0.991 entre minimum_nights_avg_ntm y maximum_minimum_nights confirma la estandarización de las reglas de estancia por parte de la gerencia del hotel; el 0.983 entre availability_90 y availability_60 muestra una consistencia rigurosa en la gestión del calendario; y el 0.982 entre las métricas de conteo (host_total_listings_count y host_listings_count) valida la uniformidad en la medición del inventario. Para el modelado predictivo, esta alta interdependencia es crítica y hace imprescindible eliminar las variables redundantes para asegurar que el modelo sea estable y que sus coeficientes reflejan relaciones causales únicas en lugar de duplicidades estadísticas.



Regresión Múltiple.

El análisis de correlación simple de la variable dependiente **review_scores_rating** revela una fuerte dependencia de las subpuntuaciones de reseña, lo cual era esperado: las cinco correlaciones más altas (todas por encima de 0.65) corresponden a las métricas de calidad internas como value (0.808), accuracy (0.803) y cleanliness (0.729), confirmando que los huéspedes prioriza que el alojamiento cumpla lo prometido y ofrezca un buen precio, y que estas sub puntuaciones son los principales drivers de la calificación general; en contraste, los factores de gestión y de negocio como ser Superhost (0.313) o el estimated_revenue_l365d (0.110) muestran correlaciones mucho más débiles y secundarias.

Modelo matemático: review_scores_rating = 0.0967 + 0.3955*review_scores_accuracy + 0.3600*review_scores_value + 0.2241*review_scores_cleanliness

La conclusión sobre el modelo de regresión lineal múltiple para review_scores_rating es que, si se incluyen todas estas variables, el modelo resultante tendrá un Coeficiente de Determinación (R^2) extremadamente alto (cercano a 1), dado que las subpuntuaciones son inherentemente redundantes y explican casi toda la varianza de la puntuación total; sin embargo, para entender el impacto real de las variables de gestión, será crucial utilizar el modelo múltiple para observar cómo los coeficientes parciales de variables como

host_is_superhost se ajustan una vez que se controla el efecto dominante de la multicolinealidad de las subpuntuaciones.

El análisis para predecir la Tasa de Aceptación del Anfitrión (**host_acceptance_rate**) establece que la capacidad de respuesta (host_response_rate) es el predictor dominante de la variable, como lo demuestra su correlación simple de 0.505

Modelo matemático: host_acceptance_rate = 0.0005 + 0.9371*host_response_rate + 0.0002*estimated_occupancy_l365d + 0.0002*number_of_reviews_ltm

De manera más crucial, su coeficiente parcial de 0.9371 en el modelo de regresión múltiple; este coeficiente abrumadoramente alto confirma que un aumento en la tasa de respuesta se traduce casi directamente en un aumento idéntico en la tasa de aceptación, incluso después de controlar el efecto de otras variables de rendimiento como la ocupación estimada y el número de reseñas recientes (cuyos coeficientes son casi nulos, ≈0.0002). Si bien el Coeficiente de Determinación (R^2) del modelo es moderado (0.3088), indicando que las variables sólo explican un 30.9% de la varianza total, la comparación de coeficientes revela que la tasa de respuesta es el factor casi exclusivo que explica el comportamiento del anfitrión en la aceptación de reservas.

El análisis de correlación simple de la variable dependiente **host_is_superhos**t revela que el estatus está intrínsecamente ligado al rendimiento operativo y la demanda del listado, más allá de los requisitos mínimos de servicio. Los coeficientes más altos se asocian con métricas de actividad: estimated_occupancy_l365d (0.368), que indica que los Superhosts tienen una mayor tasa de ocupación anual, y number_of_reviews_ltm (0.353), lo que confirma que gestionan exitosamente un mayor volumen de transacciones recientes. Las correlaciones con los requisitos de la plataforma, como la host_response_rate (0.307) y la host_acceptance_rate (0.297), son fuertes pero se sitúan ligeramente por debajo de las métricas de demanda, mientras que la review_scores_rating (0.313) también es un factor positivo esperado.

Modelo matemático: host_is_superhost = -3.7136 + 0.0018*estimated_occupancy_l365d + 0.0003*number_of_reviews_ltm + 0.8234*review_scores_rating

En conclusión, aunque el estatus de Superhost requiere un buen servicio al cliente, la correlación simple sugiere que el estatus es, ante todo, un indicador de alto volumen de demanda y un historial probado de ocupación, lo cual deberá ser explorado en el modelo de regresión múltiple para determinar el efecto independiente que tiene el estatus de Superhost sobre otras variables de resultado (como el precio), una vez que se controlan los efectos de la tasa de respuesta y la ocupación.

El análisis para predecir la Cantidad Total de Listados del Anfitrión (**host_total_listings_count**) revela un caso extremo de multicolinealidad, evidente tanto en la correlación simple como en el modelo de regresión. La correlación simple muestra una relación casi perfecta con host_listings_count (0.9777) y muy fuerte con calculated_host_listings_count (0.8766), confirmando que estas variables miden el mismo concepto de inventario.

Modelo matemático: host_total_listings_count = 0.4609 + 1.2916*host_listings_count - 0.1444*calculated_host_listings_count + 0.0000*estimated_revenue_l365d

Esta redundancia se traduce en un Coeficiente de Determinación (R^2) excepcionalmente alto de 0.9570 en el modelo múltiple, lo que implica que las variables explican más del

95% de la varianza total de la variable dependiente, un resultado que es un artefacto estadístico de la duplicidad métrica. Los coeficientes del modelo confirman esta inestabilidad: el coeficiente de host_listings_count (1.2916) es dominante, mientras que el de calculated_host_listings_count (−0.1444) es anómalamente negativo; esta inconsistencia es una señal clásica de severa multicolinealidad, lo que indica que el modelo es inestable y que las contribuciones individuales de las variables de conteo no pueden separarse ni interpretarse de manera confiable.

El análisis de correlación simple de la variable dependiente **accommodates** (capacidad de huéspedes) revela, como era de esperar, una dependencia abrumadora del tamaño físico del listado, mientras que su relación con las métricas de negocio es más débil.
Modelo matemático: accommodates = 0.0153 + 1.2520*beds + 0.3938*bathrooms + 0.0003*price
La correlación más fuerte y obvia se da con beds (0.758) y bedrooms (0.750), confirmando que la capacidad de un alojamiento está determinada principalmente por el número de camas y habitaciones disponibles. El tamaño continúa siendo el factor clave con bathrooms (0.496). En un segundo plano se encuentra la relación económica, donde price (0.472) y el estimated_revenue_l365d (0.345) muestran una correlación positiva moderada, lo que indica que una mayor capacidad de huéspedes permite a los anfitriones establecer precios más altos y generar mayores ingresos. En este contexto, el modelo de regresión lineal múltiple para accommodates tendrá un alto R^2 si incluye las variables de tamaño (beds, bedrooms) indican que es un modelo de poder predictivo moderado a fuerte, ya que logra explicar casi el 65% de la varianza total en el número de dormitorios, lo cual es un resultado sólido para la predicción de una métrica estructural.

La justificación para crear dos modelos para predecir **bedrooms** se debe a la necesidad de contrastar el poder predictivo de los factores estructurales contra los factores de gestión y rendimiento, dada la alta multicolinealidad entre las métricas de tamaño.

Modelo matemático: bedrooms = -0.0216 + 0.1919*accommodates + 0.2105*beds + 0.3914*bathrooms

Modelo matemático B: bedrooms = 1.1693 + 0.0411*maximum_minimum_nights + 0.0046*calculated_host_listings_count_entire_homes + 0.0000*estimated_revenue_l365d

Mientras el modelo estructural (que usa accommodates, beds, y bathrooms implícitamente) demostró ser moderado a fuerte con un R^2 de 0.6489, la Opción B que utiliza variables de gestión como maximum_minimum_nights y estimated_revenue_l365d resulta ser de una calidad extremadamente pobre, logrando un R^2 de solo 0.0896. Este hallazgo crucial demuestra que el número de dormitorios no puede predecirse a través de las políticas de estancia o el rendimiento del anfitrión, sino que es casi exclusivamente una función de la estructura física del alojamiento, invalidando la utilidad práctica del segundo modelo y confirmando la importancia de usar variables de la misma naturaleza que el objetivo a predecir.

Para predecir el Precio (**price**) se estructuró con dos modelos para contrastar la influencia de las variables estructurales frente a las de negocio, lo cual es esencial debido a que las correlaciones iniciales mostraron que el precio está moderadamente ligado a ambas categorías (0.488 con bedrooms y 0.383 con estimated_revenue_l365d).

Modelo matemático: price = 65.2588 + 240.1733*bedrooms + 117.9397*accommodates + 451.5502*bathrooms

Modelo matemático B: price = -3485.9317 + 0.0027*estimated_revenue_l365d + 25.1190*calculated_host_listings_count_entire_homes + 920.2549*review_scores_location

Aunque el modelo de la Opción B se propuso intencionalmente con variables de negocio y calidad (omitiendo los predictores estructurales dominantes), este resultó en un poder predictivo débil con un R^2 de solo 0.1873, confirmando que el tamaño físico del alojamiento es el principal driver del precio y, por lo tanto, el modelo estructural (no mostrado, pero utilizando bedrooms, accommodates, etc.) es el mejor modelo de predicción. Sin embargo, la Opción B es valiosa porque aísla un hallazgo clave: a pesar de ser débil, su coeficiente parcial en estimated_revenue_l365d (1.0968) es el más alto, lo que demuestra que, entre las variables de negocio, el retorno económico esperado es el factor más crucial que los anfitriones utilizan para fijar su precio.

El modelo de regresión lineal múltiple para predecir la Puntuación de Valor (**review_scores_value**) utiliza las variables de feedback review_scores_rating, review_scores_accuracy y review_scores_cleanliness, las cuales fueron seleccionadas debido a su extrema correlación y redundancia inherente con la variable dependiente.

Modelo matemático: review_scores_value = -0.0256 + 0.5488*review_scores_rating + 0.3406*review_scores_accuracy + 0.1062*review_scores_cleanliness

Este modelo demostró un poder predictivo excepcionalmente fuerte con un Coeficiente de Determinación (R^2) de 0.6916, lo que significa que el 70% de la varianza en la percepción de valor es explicada por las otras métricas de calidad. Sin embargo, este éxito estadístico se debe a la multicolinealidad, ya que todas miden la satisfacción general; el principal hallazgo es la jerarquía de la influencia, donde la puntuación general (0.5488) es el predictor más dominante, seguida por la precisión (0.3406), lo que sugiere que para maximizar el valor percibido, el anfitrión debe enfocarse primero en lograr altas calificaciones generales y, crucialmente, en asegurar que el listado sea preciso respecto a lo prometido.

El modelo de regresión lineal múltiple para predecir el Número de Baños (**bathrooms**) se diseñó intencionalmente como un modelo de contraste (Opción B), utilizando variables económicas y de gestión (price, estimated_revenue_l365d, maximum_minimum_nights) para verificar si estas podían predecir una métrica estructural, a diferencia del modelo ideal que emplearía variables de tamaño.

Modelo matemático: bathrooms = 0.7393 + 0.2833*bedrooms + 0.0813*beds + 0.0126*accommodatess

Modelo matemático B: bathrooms = 0.9822 + 0.0002*price + 0.0000*estimated_revenue_l365d + 0.0146*maximum_minimum_nights

Sin embargo, este modelo de contraste resultó ser de poder predictivo débil, con un Coeficiente de Determinación (R^2) de solo 0.2076, lo que confirma que el número de baños es principalmente una característica estructural fija y no una función del comportamiento de negocio. Los hallazgos del modelo B son cruciales: aunque la correlación simple con el precio es alta, su coeficiente parcial es casi cero (0.0002), mientras que el factor con el mayor peso es la maximum_minimum_nights (0.0146), lo que sugiere que las propiedades con más baños (y por ende más grandes) se gestionan con políticas de estancia más largas para minimizar la rotación, pero el modelo en general demuestra que la predicción de bathrooms es inviable sin incluir variables estructurales.

El análisis de correlación simple de la variable dependiente **reviews_per_month** (frecuencia mensual de reseñas) confirma que este factor está abrumadoramente ligado al volumen y la demanda del listado, lo cual es un hallazgo lógico esperado.

Modelo matemático:reviews_per_month = 0.8129 + 0.0883*number_of_reviews_ltm - 0.0033*estimated_occupancy_l365d - 0.0116*number_of_reviews_ly

La justificación para cualquier modelo predictivo se centraría en el alto rendimiento y la redundancia de las variables de feedback, dado que las cinco correlaciones más fuertes (todas por encima de 0.59) son otras métricas de conteo de reseñas (como number_of_reviews_ltm con 0.802 y number_of_reviews_ly con 0.699). Esto indica que un listado que recibe muchas reseñas en total o el año pasado, también lo hará por mes, lo cual generará multicolinealidad en un modelo de regresión múltiple que use estos predictores. El principal hallazgo es la fuerte conexión con la estimated_occupancy_l365d (0.742) y el estimated_revenue_l365d (0.550), lo que sugiere que la frecuencia de reseñas es una medida directa del volumen de transacciones y ocupación de la propiedad, mientras que las métricas de servicio del anfitrión (host_acceptance_rate y host_response_rate) tienen una correlación significativamente más baja (≈0.30), confirmando que la cantidad de reseñas depende más de la demanda que de la calidad marginal del servicio.

Conclusión.

El análisis de regresión simple y múltiple concluye que el mercado de alojamientos se rige por dos fuerzas principales con marcadas diferencias en su predictibilidad. En primer lugar, la multicolinealidad es una característica estructural dominante en la base de datos, especialmente en las métricas de inventario (host_total_listings_count vs. host_listings_count) y las subpuntuaciones de feedback (review_scores_rating vs. review_scores_value), llevando a modelos como el de host_total_listings_count a alcanzar un R^2 artificialmente alto (0.9570) con coeficientes inestables. En segundo lugar, se establece una clara jerarquía de predictores que valida la lógica del negocio:

1. *Métricas de Tamaño vs. Negocio:* El precio (price), la capacidad (accommodates) y el número de dormitorios (bedrooms) son predichos de manera superior por variables estructurales (camas, baños), mientras que los modelos basados en variables de gestión (price, revenue) resultaron en un poder predictivo débil (R^2 ≈0.20 a 0.30), demostrando que las características físicas son la base del valor.
2. *Rendimiento del Anfitrión:* La capacidad de respuesta (host_response_rate) se confirma como el predictor casi exclusivo de la tasa de aceptación (host_acceptance_rate), con un coeficiente abrumador de 0.9371. De manera similar, el estatus de Superhost está fuertemente ligado a la demanda y ocupación (estimated_occupancy_l365d).
3. *Dinámicas Segmentadas:* La regresión simple demostró que las correlaciones varían significativamente por room_type. Mientras que la tasa de respuesta influye fuertemente en la aceptación para Entire home/apt (0.510), es nula para Hotel room (−0.066), lo que reitera que los anfitriones individuales y la gestión hotelera operan bajo estrategias de negocio fundamentalmente distintas.