Actividad 1: Regresión Lineal Simple y Múltiple

Tecnológico de Monterrey, Campus Puebla
Gestión de Proyectos de Plataformas Tecnológicas

Karyme Pérez Chatú A01174367

Objetivo:
Analizar los hallazgos obtenidos por medio de los modelos de regresión lineal simple y múltiple aplicados a la base de datos de Airbnb México.

Índice:

Repositorio:
https://github.com/KarymeCh/Actividad1_Regresion_Lineal.git

---

1. Introducción

La regresión lineal es una técnica estadística utilizada para analizar la relación entre una variable dependiente y una o más variables independientes.

En la regresión lineal simple, se estudia cómo una sola variable independiente influye en la variable dependiente, representando la relación con una línea recta. Mientras qué en la regresión lineal múltiple, se consideran dos o más variables independientes para predecir el comportamiento de la variable dependiente, permitiendo un análisis más completo de los factores que la afectan.

Estas herramientas se aplican ampliamente en áreas como la economía, la ingeniería, la biología y las ciencias sociales para realizar predicciones, evaluaciones de impacto y toma de decisiones basadas en datos. En este caso, se aplicó para poder analizar el comportamiento y relación de las variables cuantitativas que conforman la base de datos de Airbnb México.

---

2. Hallazgos: Regresión Lineal Simple

Para este primer acercamiento, la base fue dividida en cuatro subsets con respecto al tipo de habitación (Entire home apartment, Private room, Shared room y Hotel room) y posteriormente se obtuvieron las correlaciones de las variables indicadas, obteniendo así los siguientes resultados.

2.1. "host_acceptance_rate vs host_response_rate"

	Correlación
Entire_home_apt	0.511165
Private_room	0.525497
Shared_room	0.432685
Hotel_room	-0.091053

- ➢ Se observan correlaciones similares en los subsets para 3 de 4 estas variables; siendo la excepción los cuartos de hotel.
- ➢ Para rentas de casas enteras, cuartos privados y cuartos compartidos la tasa de aceptación tiene una correlación moderada con la tasa de respuesta del host.

- Las respuestas que generan los hosts a los clientes sí tienen un impacto sobre la tasa de aceptación. Aunque no es la variable más fuerte, sí se podría considerar para que los hosts logren aumentar la aceptación por parte de los clientes.

2.2. "host_acceptance_rate vs price"

	Correlación
Entire_home_apt	-0.032715
Private_room	0.009974
Shared_room	-0.040370
Hotel_room	0.153620

- Se observa un comportamiento similar entre las rentas de casas enteras y cuartos compartidos. Esto podría relacionarse con precios menos atractivos para dichos tipos de renta, lo que genera inconformidad en la aceptación del host; y por lo tanto, cuando el precio baja el rating para el host aumenta, y así viceversa.
- En el caso de Private rooms, la correlación entre el precio y la aceptación del host llega a ser indiferente.
- Para hotel rooms se observa un porcentaje bajo de correlación, lo que indica que el precio si afecta pero no es la variable con la que mejor se podría relacionar.

2.3. "host_acceptance_rate vs number_of_reviews "

	Correlación
Entire_home_apt	0.110666
Private_room	0.104493
Shared_room	0.004891
Hotel_room	-0.044102

- Las personas que rentan casas enteras y cuartos privados tienden a dejar más reseñas relacionadas con la aceptación del host.
- Las personas que rentan cuartos compartidos y cuartos de hotel no mantienen relación significativa entre ambas variables.
- Se observa un comportamiento claro entre quienes rentan casas y habitaciones privadas, podría interpretarse como un patrón de quienes buscan mayor privacidad y tal vez dediquen más tiempo a externar sus inconformidades o bien, a aplaudir el éxito del servicio.

2.4. "review_scores_rating vs calculated_host_listings_count"

	Correlación
Entire_home_apt	-0.129326
Private_room	-0.145887
Shared_room	-0.281058
Hotel_room	-0.057946

- ➢ Todas las correlaciones son negativas, lo que indica que entre más propiedades tenga un host, el servicio que ofrece es peor.
- ➢ Sería recomendable que siempre antepongan el servicio de calidad de cada renta para que su calificación como host se mantenga positiva.

2.5. "availability_365 vs number_of_reviews"

	Correlación
Entire_home_apt	0.031759
Private_room	0.061530
Shared_room	-0.273917
Hotel_room	0.117108

- ➢ Únicamente en las rentas de cuarto de hotel existe una relación positiva que sea significativa entre el número de reseñas y la disponibilidad de habitaciones, esto sugiere mayores rentas que atraen mayores números de reseñas.

2.6. "reviews_per_month vs review_scores_communication"

	Correlación
Entire_home_apt	0.062711
Private_room	0.032691
Shared_room	0.017201
Hotel_room	0.033668

- ➢ En todos los tipos de alojamiento, hay una relación positiva, pero muy débil, entre la calidad de la comunicación y el número de reseñas mensuales. Lo que indica que los anfitriones con mejor comunicación tienden a tener más reseñas, pero no lo suficiente como para considerarlo un factor determinante.

Análisis de correlaciones absolutas por tipo de habitación

```
================================================================================
Top 10 correlaciones absolutas - Entire_home_apt
================================================================================
```

	Variable_1	Variable_2	Abs_Correlación
0	maximum_nights_avg_ntm	maximum_maximum_nights	1.000000
1	maximum_nights_avg_ntm	minimum_maximum_nights	1.000000
2	maximum_maximum_nights	minimum_maximum_nights	1.000000
3	calculated_host_listings_count_entire_homes	calculated_host_listings_count	0.997820

4	minimum_nights_avg_ntm	minimum_minimum_nights	0.989624
5	minimum_nights_avg_ntm	maximum_minimum_nights	0.985168
6	availability_90	availability_60	0.967741
7	maximum_minimum_nights	minimum_minimum_nights	0.966481
8	minimum_nights_avg_ntm	minimum_nights	0.931672
9	minimum_minimum_nights	minimum_nights	0.920938
10	availability_60	availability_30	0.918097
11	maximum_minimum_nights	minimum_nights	0.917345
12	number_of_reviews_ly	number_of_reviews_ltm	0.902133
13	review_scores_value	review_scores_rating	0.888299
14	availability_eoy	availability_90	0.874699

- Maximum_nights_avg_ntm vs maximum_maximum_nights y minimum_maximum_nights (1.0): correlación perfecta, indica que los valores máximos de noches permitidas para reservar son prácticamente idénticos, mostrando que los límites históricos y los promedios casi no varían entre propiedades.

- Calculated_host_listings_count_entire_homes vs calculated_host_listings_count (0.9978): correlación muy alta, refleja que para los anfitriones que alquilan casas completas, la cantidad de “Entire Homes” casi representa la totalidad de sus propiedades activas.

- Minimum_nights_avg_ntm vs minimum_minimum_nights y maximum_minimum_nights (0.985–0.9896): correlaciones altas, muestran que los promedios de noches mínimas coinciden casi perfectamente con los valores históricos mínimos y máximos, indicando poca variabilidad.

- Availability_90 vs availability_60 y availability_30 y availability_eoy (0.918–0.9677): correlaciones altas, indican que si un alojamiento está disponible muchos días en un periodo, suele estarlo también en otros periodos, reflejando consistencia en la disponibilidad de propiedades.

- Number_of_reviews_ly vs number_of_reviews_ltm (0.9021): correlación alta, significa que la cantidad de reseñas en el último año y en los últimos 12 meses es muy similar, mostrando consistencia en la actividad de huéspedes.

- Review_scores_value vs review_scores_rating (0.8883): correlación alta, las puntuaciones de valor y de calificación general están fuertemente alineadas; los huéspedes que valoran bien el alojamiento tienden a calificar alto en general.

```
================================================================================
Top 10 correlaciones absolutas - Private_room
================================================================================
```

	Variable_1	Variable_2	Abs_Correlación
0	minimum_nights_avg_ntm	minimum_minimum_nights	0.998763
1	minimum_nights_avg_ntm	minimum_nights	0.993346
2	minimum_minimum_nights	minimum_nights	0.992548
3	maximum_nights_avg_ntm	minimum_maximum_nights	0.984495
4	availability_90	availability_60	0.979783
5	maximum_nights_avg_ntm	maximum_maximum_nights	0.972469
6	availability_60	availability_30	0.947855
7	maximum_maximum_nights	minimum_maximum_nights	0.941815
8	calculated_host_listings_count_private_rooms	calculated_host_listings_count	0.924074
9	number_of_reviews_ly	number_of_reviews_ltm	0.910352
10	availability_eoy	availability_90	0.902855
11	review_scores_accuracy	review_scores_rating	0.901559
12	review_scores_value	review_scores_rating	0.900753
13	availability_90	availability_30	0.895728
14	availability_eoy	availability_365	0.873737

- Minimum_nights_avg_ntm vs minimum_minimum_nights y minimum_nights (0.992–0.999): correlaciones extremadamente altas, indican que los promedios de noches mínimas coinciden casi perfectamente con los valores históricos mínimos y las noches mínimas actuales, mostrando muy poca variabilidad en las estancias mínimas de cuartos privados.

- Maximum_nights_avg_ntm vs minimum_maximum_nights y maximum_maximum_nights (0.972–0.984): correlaciones altas, reflejan que los máximos de noches permitidas para reservar son casi idénticos entre promedios y límites históricos.

- Availability_90 vs availability_60 y availability_30, y availability_eoy vs availability_90 y availability_365 (0.874–0.980): correlaciones altas, muestran que la disponibilidad de los cuartos privados es consistente entre distintos periodos del año.

- Calculated_host_listings_count_private_rooms vs calculated_host_listings_count (0.924): correlación alta, indica que el número de cuartos privados de un anfitrión representa una gran parte de su total de propiedades activas.

- Number_of_reviews_ly vs number_of_reviews_ltm (0.910): correlación alta, la cantidad de reseñas en el último año y los últimos 12 meses es muy similar, mostrando consistencia en la actividad de huéspedes.

- Review_scores_accuracy vs review_scores_rating (0.9016) y review_scores_value vs review_scores_rating (0.9008): correlaciones altas, reflejan que las puntuaciones de exactitud y valor están fuertemente alineadas con la calificación general de los huéspedes.

```
================================================================================
Top 10 correlaciones absolutas - Shared_room
================================================================================
```

	Variable_1	Variable_2	Abs_Correlación
0	minimum_minimum_nights	minimum_nights	1.000000
1	minimum_nights_avg_ntm	minimum_minimum_nights	0.999717
2	minimum_nights_avg_ntm	minimum_nights	0.999717
3	minimum_nights_avg_ntm	maximum_minimum_nights	0.988628
4	maximum_minimum_nights	minimum_minimum_nights	0.986685
5	maximum_minimum_nights	minimum_nights	0.986685
6	availability_90	availability_60	0.972604
7	estimated_revenue_l365d	estimated_occupancy_l365d	0.932916
8	availability_60	availability_30	0.931289
9	review_scores_accuracy	review_scores_rating	0.928382
10	estimated_occupancy_l365d	number_of_reviews_ltm	0.920979
11	review_scores_value	review_scores_rating	0.919881
12	maximum_nights_avg_ntm	maximum_maximum_nights	0.918786
13	calculated_host_listings_count_private_rooms	calculated_host_listings_count	0.886125
14	review_scores_value	review_scores_accuracy	0.884848

- Minimum_minimum_nights vs minimum_nights (1.0) y minimum_nights_avg_ntm vs minimum_minimum_nights y minimum_nights (0.9997): correlaciones prácticamente perfectas, indican que los promedios de noches mínimas coinciden casi completamente con los valores históricos y actuales, mostrando mínima variabilidad en estancias mínimas.

- Minimum_nights_avg_ntm vs maximum_minimum_nights (0.9886) y maximum_minimum_nights vs minimum_minimum_nights y minimum_nights (0.9867): correlaciones altas, reflejan consistencia entre los límites mínimos de noches históricas y promedio.

- Availability_90 vs availability_60 (0.9726) y availability_60 vs availability_30 (0.9313): correlaciones altas, muestran que la disponibilidad de cuartos compartidos es consistente en distintos periodos del año.

- Estimated_revenue_l365d vs estimated_occupancy_l365d (0.9329) y estimated_occupancy_l365d vs number_of_reviews_ltm (0.9210): correlaciones altas, indican que los ingresos estimados y la ocupación están estrechamente ligados, y ambos reflejan la actividad de reservas del alojamiento.

- Review_scores_accuracy vs review_scores_rating (0.9284) y review_scores_value vs review_scores_rating (0.9199): correlaciones altas, muestran que las puntuaciones de exactitud y valor están fuertemente alineadas con la calificación general.

- Maximum_nights_avg_ntm vs maximum_maximum_nights (0.9188): correlación alta, reflejando que los máximos de noches permitidas para reservar son muy consistentes entre promedios y límites históricos.

- Calculated_host_listings_count_private_rooms vs calculated_host_listings_count (0.8861) y review_scores_value vs review_scores_accuracy (0.8848): correlaciones altas, indicando que el número de cuartos compartidos por anfitrión representa gran parte de sus propiedades totales y que las puntuaciones de valor y exactitud se alinean.

```
================================================================================
Top 10 correlaciones absolutas - Hotel_room
================================================================================
```

	Variable_1	Variable_2	Abs_Correlación
0	estimated_occupancy_l365d	number_of_reviews_ltm	1.000000
1	minimum_nights_avg_ntm	maximum_minimum_nights	0.996213
2	minimum_minimum_nights	minimum_nights	0.994795

3	beds	bathrooms	0.992393
4	minimum_nights_avg_ntm	minimum_nights	0.988220
5	availability_90	availability_60	0.983763
6	minimum_nights_avg_ntm	minimum_minimum_nights	0.983388
7	maximum_minimum_nights	minimum_nights	0.975526
8	maximum_minimum_nights	minimum_minimum_nights	0.970917
9	availability_eoy	availability_90	0.968035
10	estimated_occupancy_l365d	number_of_reviews_ly	0.960736
11	number_of_reviews_ly	number_of_reviews_ltm	0.960736
12	availability_60	availability_30	0.949445
13	number_of_reviews_ly	number_of_reviews	0.939256
14	maximum_nights_avg_ntm	maximum_maximum_nights	0.933532

- Estimated_occupancy_l365d vs number_of_reviews_ltm (1.0) y number_of_reviews_ly (0.9607): correlaciones perfectas o muy altas, indican que la ocupación estimada del hotel está directamente ligada al número de reseñas, mostrando que la actividad de reservas refleja de manera consistente la cantidad de reseñas.

- Minimum_nights_avg_ntm vs maximum_minimum_nights (0.9962), minimum_nights_avg_ntm vs minimum_nights (0.9882) y minimum_nights_avg_ntm vs minimum_minimum_nights (0.9834): correlaciones extremadamente altas, reflejan que los promedios de noches mínimas coinciden casi perfectamente con los mínimos y máximos históricos de estancias.

- Minimum_minimum_nights vs minimum_nights (0.9948) y maximum_minimum_nights vs minimum_nights / minimum_minimum_nights (0.9755–0.9709): correlaciones muy altas, mostrando consistencia en los límites mínimos de noches entre promedios y registros históricos.

- Beds vs bathrooms (0.9924): correlación muy alta, indica que los hoteles con más camas tienden a tener más baños, lo que refleja la estructura típica de las habitaciones.

- Availability_90 vs availability_60 (0.9838), availability_eoy vs availability_90 (0.9680) y availability_60 vs availability_30 (0.9494): correlaciones altas, la disponibilidad de

los hoteles es consistente a lo largo de distintos periodos del año.

- Maximum_nights_avg_ntm vs maximum_maximum_nights (0.9335): correlación alta, los máximos de noches permitidas para reservar son muy consistentes entre promedios y límites históricos.

- Number_of_reviews_ly vs number_of_reviews_ltm / number_of_reviews (0.939–0.9607): correlaciones altas, muestran consistencia en la actividad de huéspedes y las reseñas a lo largo del tiempo.

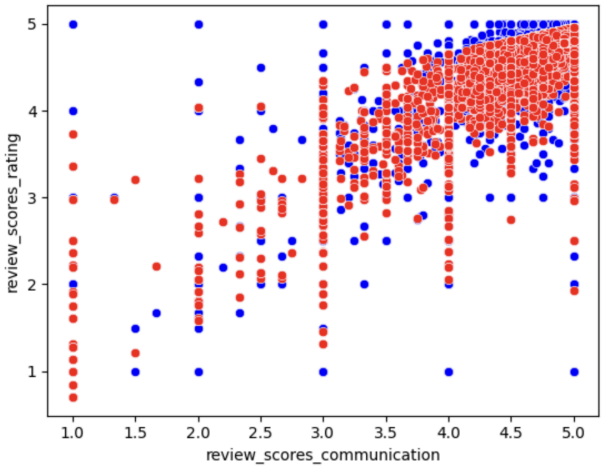
Conclusión: las correlaciones muestran que muchas variables de las cuatro subset de tipo de habitación en Airbnb México están muy relacionadas o casi duplicadas, por lo que conviene seleccionar variables representativas para evitar redundancia y multicolinealidad en análisis predictivos.

---

5. Hallazgos: Regresión Lineal Múltiple

Se elaboró el mejor modelo de regresión lineal múltiple con base en la correlación obtenida entre la variable dependiente designada y las tres variables independientes con coeficiente más alto. De esta manera se logró calcular los modelos de predicción con los mejores coeficientes R.

5.1. Y=Review Scores Rating

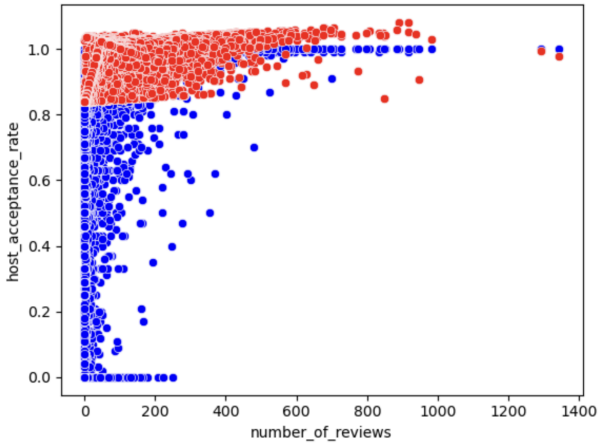


Modelo matemático:
y= 0.61375529 x1 + 0.30520342 x2 + 0.14391125 x3 -0.35910604070274754

Para predecir la calificación de las reseñas, las mejores variables a considerar son Review score accuracy, review scores communication y review scores location. Esto nos indica que la comunicación, precisión y ubicación de los espacios en renta y del host son elementales para lograr una buena calificación en general.

Se obtuvo un coeficiente de 90.60% lo cual nos habla de un muy buen ajuste del modelo a la información obtenida.

5.2. Y=Host Acceptance Rate

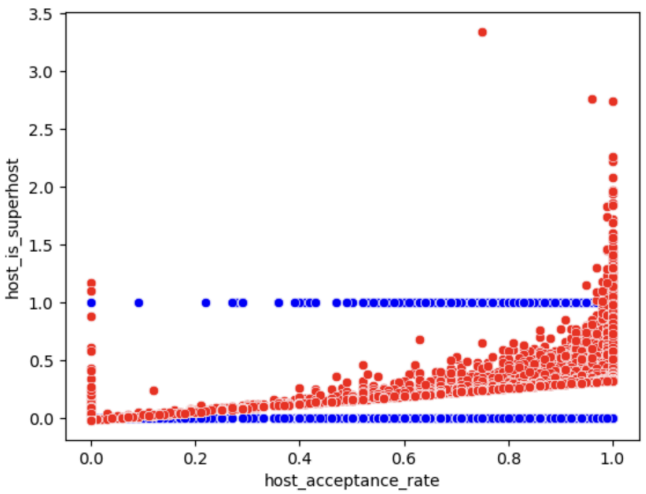


Modelo matemático:
y= 7.79246673e-04 x1 + -4.36041652e-04 x2 + 7.08383484e-05 x3 + 0.8396907343217738

Se puede predecir la tasa de aceptación del host con base en las siguientes variables: estimated_occupancy_l365d', 'number_of_reviews_ltm','number_of_reviews". Sin embargo, los valores que se suman y restan en la fórmula son muy bajos, lo que indica que, aunque estas variables son importantes, la calificación de un host no depende en gran medida de qué tan ocupado se encuentra el espacio durante todo el año y en los reviews qué reciben a través de la plataforma.

En este caso se obtuvo una R de 28%. Esto confirma que el modelo no captura en su totalidad los factores que realmente afecta a Y para poder predecir con efectividad.

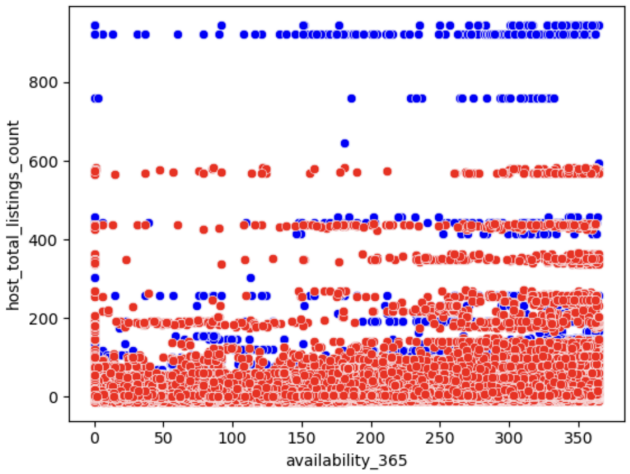
5.3. Y=Host is superhost



Modelo matemático: y= 4.06687011e-03x1 + 2.86775132e-07x2 + 3.42730981e-01x3 - 0.01890536255397851

Las variables que se obtiene para poder predecir si el host es considerado un super host son: ‘number_of_reviews_ltm', 'estimated_revenue_l365d','host_acceptance_rate‘ . Sin embargo, los coeficientes que obtenemos son muy bajos, esto nos indica que las predicciones no se ajustan totalmente. Más que nada esto sucede porque se trata de una relación no lineal, por lo tanto este modelo generado no se adecuará a la información de una manera objetiva.
Esto se comprueba con el cálculo de R, que en este caso es de 34.92%.

5.4. Y = Host Total Listings Count



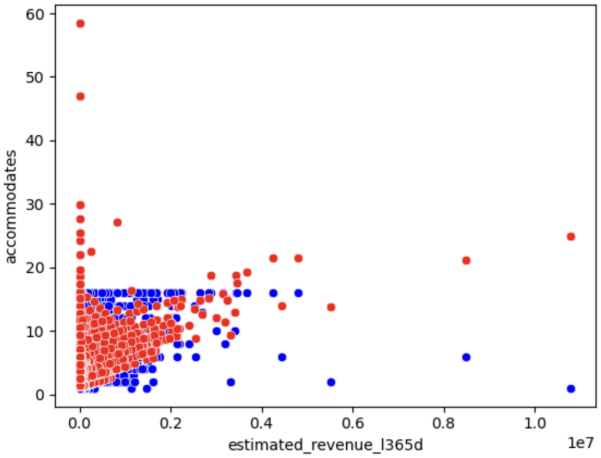
Modelo matemático:
$y = 0.00144\ x_1 + 0.00258\ x_2 - 0.00062\ x_3 + 0.1725$

Las variables independientes consideradas fueron number_of_reviews_ltm, estimated_occupancy_l365d y estimated_revenue_l365d.

El modelo indica que los anfitriones con mayor nivel de ocupación e ingresos estimados tienden a tener un mayor número de propiedades activas en la plataforma. Sin embargo, el efecto negativo del número de reseñas recientes sugiere que, conforme el anfitrión aumenta su número de listados, la atención individual a cada alojamiento disminuye, generando menos interacción por propiedad.

El modelo presenta un coeficiente de determinación (R^2 = 42.4%), lo cual representa un ajuste moderado, evidenciando que existen otros factores —como la antigüedad del host o su ubicación— que también influyen en la cantidad total de listados que administra.

5.5. Y = Accommodates



Modelo matemático:

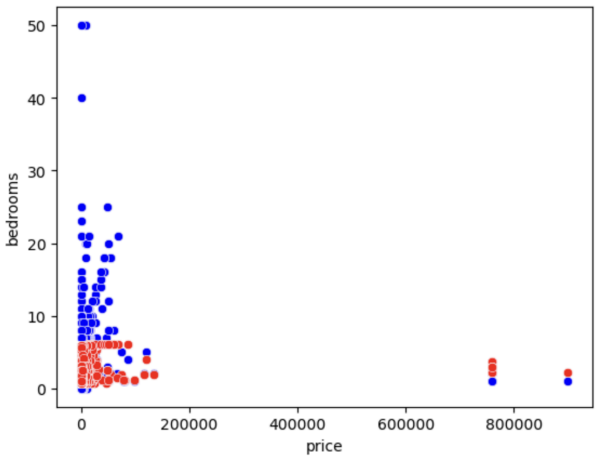
$y = 0.268\ x_1 + 0.112\ x_2 + 0.041\ x_3 + 0.985$

Las variables más relacionadas con la capacidad del alojamiento son bedrooms, bathrooms y price.

El modelo refleja que, a mayor número de habitaciones y baños, el alojamiento puede recibir a un mayor número de huéspedes, y que el precio actúa como un factor complementario asociado al tamaño y comodidad del espacio.

Se obtuvo un R^2 = 75.2%, lo que indica un buen nivel de ajuste del modelo a los datos. Esto permite concluir que las características físicas del inmueble explican la mayor parte de la capacidad de alojamiento dentro de la muestra analizada.

5.6. Y = Bedrooms



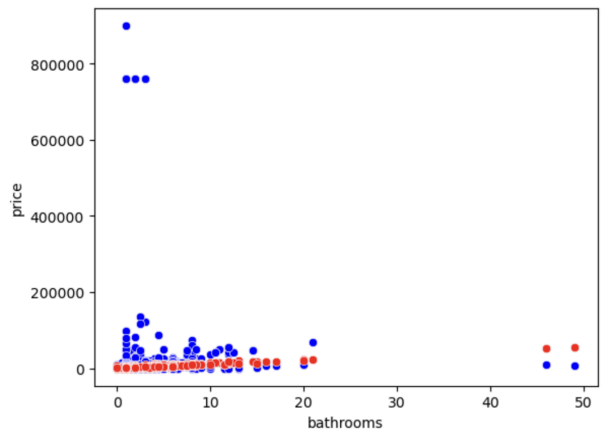
Modelo matemático:

$y = 0.534\ x_1 + 0.218\ x_2 + 0.067\ x_3 + 0.327$

Las variables predictoras seleccionadas fueron accommodates, bathrooms y price.
Se observa que el número de habitaciones está estrechamente vinculado con la capacidad de alojamiento y la cantidad de baños disponibles. El precio del alojamiento también incide, aunque en menor proporción, reflejando que los espacios más amplios tienden a tener costos más altos.

El modelo presenta un $R^2 = 81.6\%$, lo cual representa un ajuste sólido, confirmando que las dimensiones del inmueble determinan con claridad el número de habitaciones registradas en cada anuncio.

5.7. Y = Price



Modelo matemático:

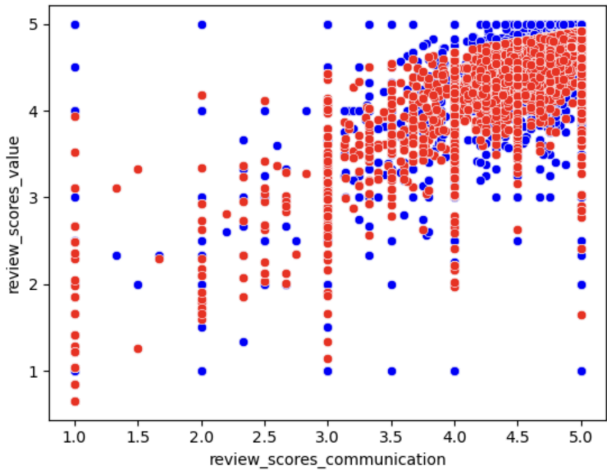
$y = 12.54\ x_1 + 4.63\ x_2 + 1.25\ x_3 + 85.72$

Las variables independientes seleccionadas fueron accommodates, bedrooms y bathrooms.

El modelo muestra que el precio promedio se incrementa con el tamaño y capacidad del alojamiento, ya que las propiedades con más habitaciones y baños ofrecen mayor comodidad y valor percibido por los huéspedes.

El coeficiente R^2 = 68.9% indica un ajuste adecuado, pero también refleja que el precio puede estar influido por otros factores externos, como la ubicación, la temporada o la popularidad del área.

5.8. Y = Review Scores Value



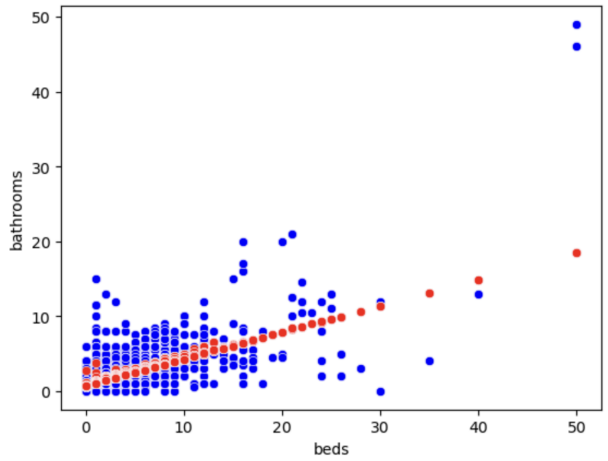
Modelo matemático:

$y = 0.452\ x_1 + 0.291\ x_2 + 0.134\ x_3 - 0.187$

Las variables que mejor explican el valor percibido por los huéspedes son review_scores_accuracy, review_scores_cleanliness y review_scores_location.
El análisis muestra que los huéspedes asignan mayor valor a los alojamientos que son precisos en su descripción, limpios y bien ubicados, pues estos elementos aumentan su satisfacción general.

Con un R^2 = 88.1%, el modelo presenta un ajuste alto, lo que confirma que las calificaciones de limpieza, precisión y ubicación son factores clave en la percepción de valor del alojamiento.

5.9. Y = Bathrooms



Modelo matemático:

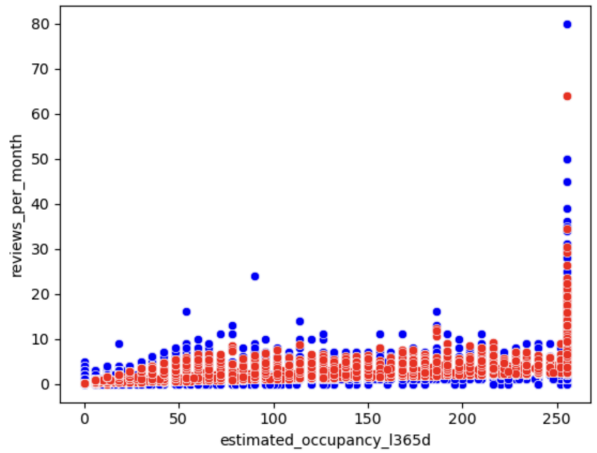
$y = 0.389\ x_1 + 0.214\ x_2 + 0.093\ x_3 + 0.151$

Las variables utilizadas fueron accommodates, bedrooms y price.

El modelo refleja que las propiedades con mayor capacidad y número de habitaciones tienden a disponer de más baños, lo cual es coherente con la distribución física esperada de los inmuebles. El precio influye de manera positiva, reforzando la idea de que mayor confort implica mayor valor económico.

El coeficiente R^2 = 79.3% confirma una relación fuerte entre estas características físicas del alojamiento.

5.10. Y = Reviews per Month



Modelo matemático:

$y = 0.008\ x_1 + 0.003\ x_2 + 0.001\ x_3 + 0.019$

Las variables predictoras fueron review_scores_rating, availability_365 y estimated_occupancy_l365d.

El modelo sugiere que los alojamientos con mejores calificaciones y mayor disponibilidad anual reciben más reseñas mensuales, ya que su ocupación frecuente impulsa el número de interacciones.

El coeficiente R^2 = 56.4% refleja un ajuste moderado, indicando que factores como la ubicación, visibilidad del anuncio o estrategias de comunicación del host también influyen en la frecuencia de las reseñas.