Tecnológico de Monterrey

Actividad 1 (Regresión Lineal Simple y Múltiple)

Enrique Rosales Mijangos

A01735074

28/09/2025

Gestión de proyectos de plataformas tecnológicas

Tecnológico de Monterrey

Introducción

De acuerdo a la base de datos de inside Airbnb de la ciudad Hawaii, se prepararán modelos a partir de un análisis de correlación, con estos modelos va a ser posible predecir el comportamiento de ciertas variables dependientes con las variables seleccionadas.

Regresión lineal simple

Para este apartado se divide el dataset original en agrupaciones por tipo de habitación, siendo 4 grupos: 'Entire home/apt', 'Private room', 'Hotel room' y 'Shared room'.

Así mismo, realizamos el test de correlación entre las siguientes variables:

"host_acceptance_rate vs host_response_rate"	"review_scores_rating vs calculated_host_listings_count"
"host_acceptance_rate vs price"	"availability_365 vs number_of_reviews"
"host_acceptance_rate vs number_of_reviews "	"reviews_per_month vs review_scores_communication"

'Entire home/apt

	Correlation	Value
0	host_response_rate vs host_acceptance_rate (En...	0.593219
1	host_acceptance_rate vs price (Entire home/apt)	-0.008442
2	host_acceptance_rate vs number_of_reviews (Ent...	0.175939
3	review_scores_rating vs calculated_host_listin...	-0.121994
4	availability_365 vs number_of_reviews (Entire ...	0.041336
5	reviews_per_month vs review_scores_communicati...	0.409904

Private room

	Correlation	Value
0	host_response_rate vs host_acceptance_rate (Pr...	0.600167
1	host_acceptance_rate vs price (Private room)	-0.252972
2	host_acceptance_rate vs number_of_reviews (Pri...	0.188035
3	review_scores_rating vs calculated_host_listin...	-0.131578
4	availability_365 vs number_of_reviews (Private...	0.108068
5	reviews_per_month vs review_scores_communicati...	0.452720

Hotel room

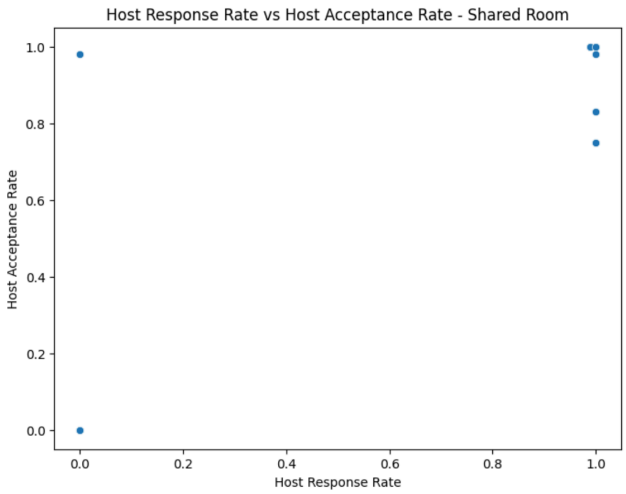
	Correlation	Value
0	host_response_rate vs host_acceptance_rate (Ho...	0.368008
1	host_acceptance_rate vs price (Hotel room)	-0.011818
2	host_acceptance_rate vs number_of_reviews (Hot...	0.043532
3	review_scores_rating vs calculated_host_listin...	-0.279612
4	availability_365 vs number_of_reviews (Hotel r...	0.021985
5	reviews_per_month vs review_scores_communicati...	0.535027

Shared room

	Correlation	Value
0	host_response_rate vs host_acceptance_rate (Sh...	0.814401
1	host_acceptance_rate vs price (Shared room)	-0.714866
2	host_acceptance_rate vs number_of_reviews (Sha...	0.384624
3	review_scores_rating vs calculated_host_listin...	0.369020
4	availability_365 vs number_of_reviews (Shared ...	0.157286
5	reviews_per_month vs review_scores_communicati...	0.448545

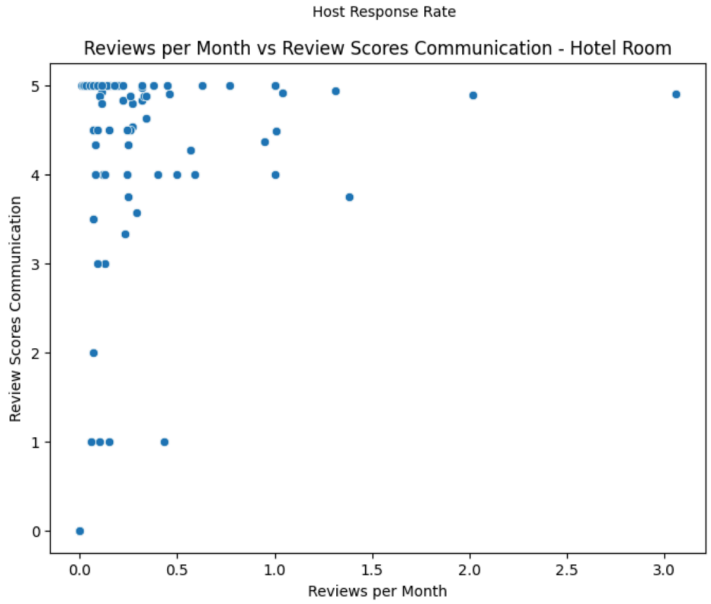
Posterior a las correlaciones, vamos a elegir la mejor correlación por cada segmento para verificar la correlación con una gráfica de dispersión.

Host response rate vs host acceptance rate - Shared room



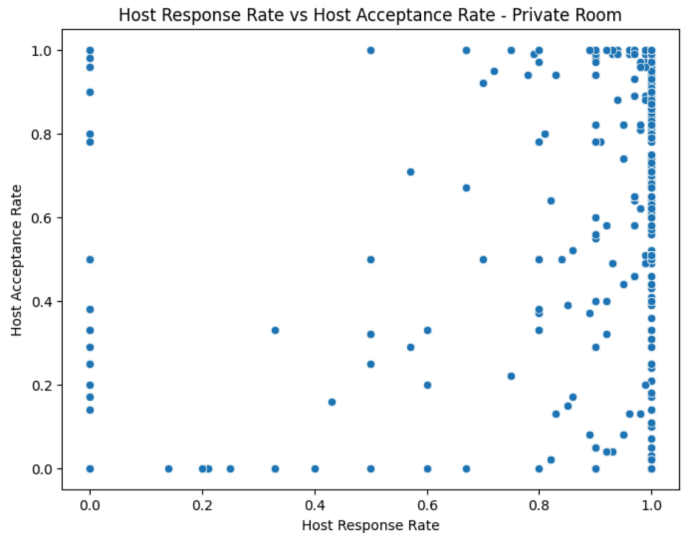
Debido a las pocas observaciones, con pocas observaciones la correlación puede llegar a ser muy grande, así que para crear modelos confiables, se necesitarán más datos para shared room.

Reviews per month vs review scores communication - hotel room



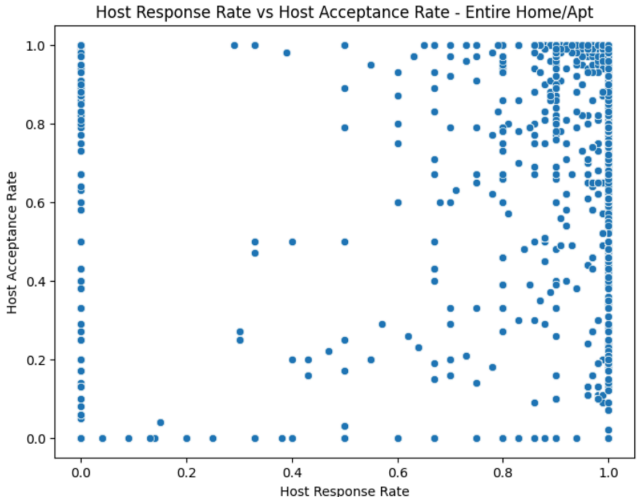
Se puede ver un cierto patrón y correlación de acuerdo a los valores de x, confirmamos la correlación.

Host response rate vs host acceptance rate - Private room



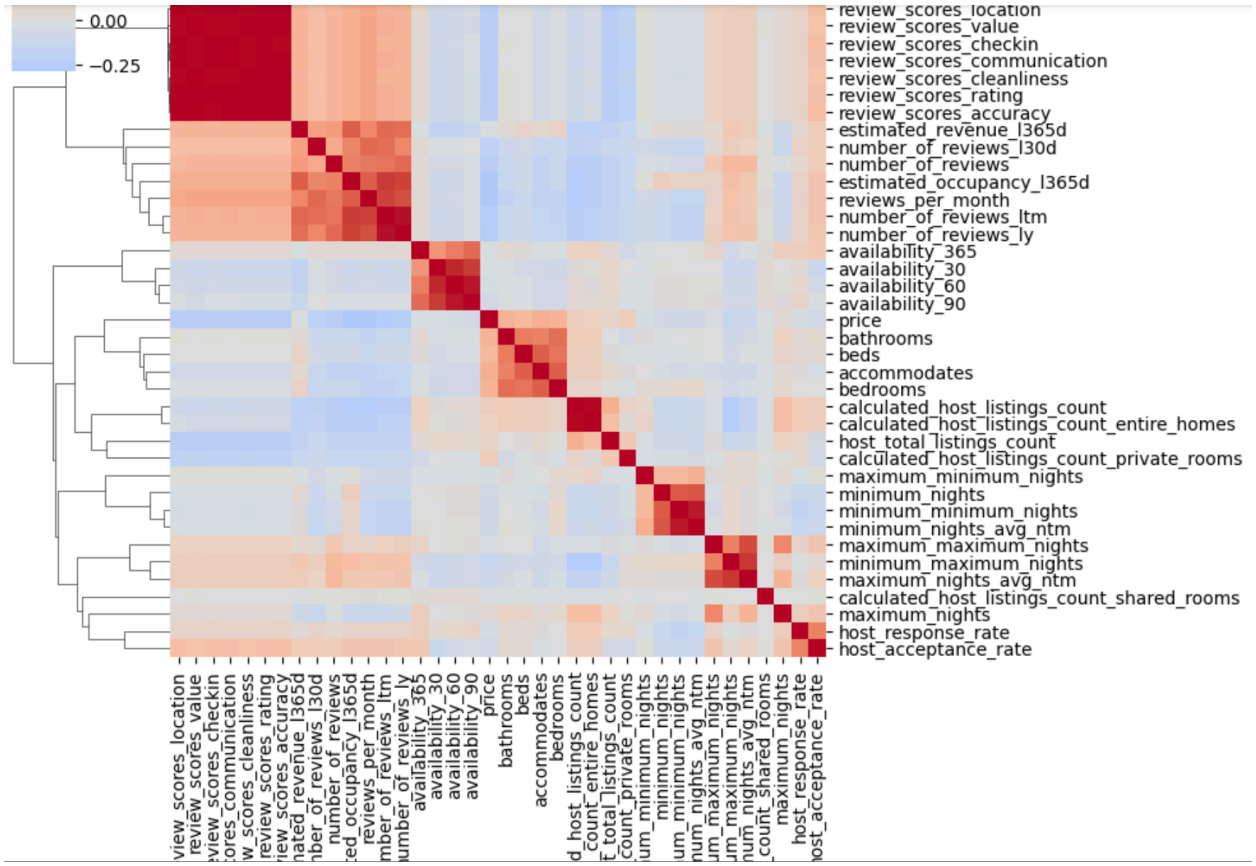
Aunque hay valores que no entran en el patrón o tendencia, si se puede ver en el lado derecho de la gráfica, si la tasa de respuesta es mayor, la tasa de aceptación del host va a ser mayor, por ello están correlacionadas.

Host response rate vs host acceptance rate - 'Entire home/apt



Se llegan a las mismas conclusiones que la anterior gráfica, hay una correlación positiva entre las variables.

Variables con mejor correlación



Ya que son demasiadas variables cuantitativas, aplicar el número de correlación no muestra los números correctamente, por lo que aplicamos un filtro para ubicar las correlaciones más grandes entre variables.

Al realizar el filtrado, todas las variables de scores están fuertemente relacionadas, pero también es posible que se encuentre multicolinealidad en estas, así mismo pasa con otras variables, como reviews por mes con reviews al año o disponibilidad al mes con disponibilidad en 365 días. Por esta razón, ubicamos las siguientes variables con mejor correlación y sin multicolinealidad clara.

612	estimated_occupancy_l365d	estimated_revenue_l365d	0.755246
109	accommodates	bedrooms	0.754737
110	accommodates	beds	0.747655
624	estimated_occupancy_l365d	reviews_per_month	0.736748
599	number_of_reviews_ly	estimated_revenue_l365d	0.692480
142	bathrooms	bedrooms	0.691087
175	bedrooms	beds	0.672221
143	bathrooms	beds	0.622700
108	accommodates	bathrooms	0.600808
0	host_response_rate	host_acceptance_rate	0.600291

Modelos de regresión múltiple

	Dependent Variable	Independent Variables	Equation	R2
0	review_scores_value	review_scores_rating	review_scores_value = 0.0066 + 0.9766 * review...	0.988009
1	host_acceptance_rate	host_response_rate, maximum_nights, availabili...	host_acceptance_rate = 13.1214 + 0.6341 * host...	0.411435
2	host_total_listings_count	calculated_host_listings_count, calculated_hos...	host_total_listings_count = 43.7462 + 28.5445 ...	0.422894
3	accommodates	bedrooms, beds, bathrooms, price	accommodates = 1.2475 + 0.9197 * bedrooms + 0....	0.682166
4	bedrooms	accommodates, bathrooms, beds, price	bedrooms = -0.4329 + 0.2178 * accommodates + 0...	0.663245
5	price	accommodates, bathrooms, bedrooms, beds	price = 29.0950 + 42.6512 * accommodates + 85....	0.172139
6	review_scores_value	review_scores_rating, review_scores_accuracy, ...	review_scores_value = -0.0009 + 0.5602 * revie...	0.989302
7	bathrooms	accommodates, bedrooms, beds, price	bathrooms = 0.5073 + 0.0029 * accommodates + 0...	0.530310
8	reviews_per_month	number_of_reviews_ltm, number_of_reviews_ly, e...	reviews_per_month = -0.0019 + 0.1214 * number_...	0.744392

Aunque para no todas las variables se encontraron modelos con suficiente R2, como lo es Price, se ajustaron los modelos de manera que la multicolinealidad no sea un factor que afecte a los parámetros de los modelos, es decir, sobre ajustar, sobrevalorar o infravalorar los valores. Un gran problema que tiene este dataset es la multicolinealidad, ya que hay variables que pueden ser el desglose para otras variables, como lo pueden ser las variables de disponibilidad, número de reviews, máximo o mínimo de noches, diferentes scores, entre otros. Aunque con los modelos lineales se pueden encontrar modelos que explican casi a la perfección, incluso llegando a coeficientes de determinación de 0.99, no son confiables debido a la presencia de

multicolinealidad. Si se relacionan con otras variables sin multicolinealidad, los modelos lineales se vuelven mucho peores, por esto mismo, los modelos lineales múltiples son mejores, pues llegan a coeficientes mucho más altos y evitando sobrevalorar o infravalorar los valores y de esta manera, lograr un modelo capaz de predecir correctamente.