Tecnológico de Monterrey

Actividad 2.1 Regresión Lineal Simple y Múltiple

Pilar Méndez Briones | A01736843

Analítica de datos y herramientas de inteligencia artificial

Alfredo García Suárez

Fecha de entrega

6 de abril de 2025

Introducción

Este análisis se realizó para identificar qué tan relacionadas están algunas variables clave en alojamientos publicados en una plataforma de hospedaje. Se evaluaron diferentes tipos de habitaciones —Entire home/apt, Private room, Shared room y Hotel room—, observando cómo se comportan ciertos indicadores en cada categoría.

Resumen del proceso

Primero, se limpió el conjunto de datos y se filtraron las columnas numéricas. Luego, se dividió la información por tipo de habitación. Para cada tipo, se realizaron regresiones lineales simples entre pares específicos de variables. Se calculó el coeficiente de determinación (R^2), que indica qué tanto una variable puede predecir a otra, y el coeficiente de correlación (r), que mide qué tan fuerte es la relación entre ambas. Los resultados se compararon para analizar el comportamiento de cada variable según el tipo de habitación.

Análisis de las variables elegidas por el profe

Tipo de habitación	Par de variables	R^2	r
Entire home/apt	availability_365 vs number_of_reviews	0.029274618	0.171098271
Hotel room	availability_365 vs number_of_reviews	0.200939114	0.448262328
Private room	availability_365 vs number_of_reviews	0.013628748	0.116742231
Shared room	availability_365 vs number_of_reviews	0.082822391	0.287788655
Entire home/apt	host_acceptance_rate vs host_response_rate	0.0	0.0
Hotel room	host_acceptance_rate vs host_response_rate	0.0	0.0
Private room	host_acceptance_rate vs host_response_rate	0.0	0.0
Shared room	host_acceptance_rate vs host_response_rate	0.0	0.0
Entire home/apt	host_acceptance_rate vs number_of_reviews	0.020681004	0.143808916
Hotel room	host_acceptance_rate vs number_of_reviews	0.210240674	0.458520091
Private room	host_acceptance_rate vs number_of_reviews	0.045469252	0.213235203
Shared room	host_acceptance_rate vs number_of_reviews	0.012854259	0.113376624
Entire home/apt	host_acceptance_rate vs price	1.023E-06	0.001011381
Hotel room	host_acceptance_rate vs price	0.246872942	0.496863103
Private room	host_acceptance_rate vs price	0.009603876	0.097999369
Shared room	host_acceptance_rate vs price	0.004542116	0.067395222
Entire home/apt	review_scores_location vs review_scores_cleanliness	0.075780503	0.275282587
Hotel room	review_scores_location vs review_scores_cleanliness	0.012084799	0.109930886

Private room	review_scores_location vs review_scores_cleanliness	0.067472867	0.259755398
Shared room	review_scores_location vs review_scores_cleanliness	0.442659803	0.665326839
Entire home/apt	reviews_per_month vs review_scores_communication	0.021788098	0.147607919
Hotel room	reviews_per_month vs review_scores_communication	0.019314127	0.138975278
Private room	reviews_per_month vs review_scores_communication	0.015912954	0.126146559
Shared room	reviews_per_month vs review_scores_communication	0.182367626	0.427045227

1. host_acceptance_rate vs host_response_rate

En todos los tipos de habitación, el valor de R^2 fue exactamente 0, lo que indica ninguna relación lineal entre la tasa de aceptación del anfitrión y su velocidad de respuesta. Esto puede deberse a que ambos indicadores están determinados por mecanismos distintos: aceptar una solicitud es una decisión humana o automática basada en disponibilidad, mientras que responder puede ser más inmediato o gestionado por bots o aplicaciones, especialmente en hoteles o cuentas múltiples.

2. host_acceptance_rate vs price

Aquí se observó una relación débil o muy débil en la mayoría de los casos. En Entire home/apt y Shared room, el R^2 fue prácticamente 0, lo que sugiere que el precio del alojamiento no depende de la disposición del anfitrión a aceptar reservaciones

3. host_acceptance_rate vs number_of_reviews

Puede esperarse que un anfitrión que acepta más reservaciones reciba más reseñas, esto no se refleja directamente en los datos.

4. review_scores_location vs review_scores_cleanliness

En Shared room, el R^2 fue de 0.44, lo que representa una relación fuerte: cuando los huéspedes valoran bien la ubicación, tienden también a dar buenas notas a la limpieza. Esto podría deberse a que alojamientos compartidos bien ubicados suelen cuidarse más o son percibidos con mayor atención.

Análisis de la tabla que envié en Github sobre los 10 coeficientes más altos

Entire home/ apt

Variables: availability_90 vs availability_60

$R^2 = 0.965$, $r = 0.983$

Esta relación indica que los hogares o departamentos completos que están disponibles en los últimos 90 días también suelen estarlo en los últimos 60.

Hotel room

Variables: availability_90 vs availability_60

$R^2 = 0.975$, $r = 0.987$

La disponibilidad en los últimos 90 días y la de los últimos 60 días están fuertemente relacionadas. Esto indica que cuando un hotel está activo en un periodo, también suele estarlo en otro cercano. En hoteles, la disponibilidad suele ser continua.y por eso es que salen estos valores.

Private room

Variables: maximum_nights_avg_ntm vs maximum_maximum_nights

$R^2 = 0.997$, $r = 0.999$

La media de noches máximas permitidas y el máximo absoluto de noches permitidas están fuertemente correlacionadas. Tiene sentido, porque el promedio (avg_ntm) está relacionado con el valor máximo.

Shared room

Variables: availability_90 vs availability_60

$R^2 = 0.977$, $r = 0.989$

La disponibilidad en los últimos 90 días y en los últimos 60 están fuertemente correlacionadas. Esto tiene sentido, ya que cuando un alojamiento compartido está activo durante un periodo largo, normalmente también lo está en uno más corto. Esta relación refleja constancia operativa por parte del anfitrión.

Análisis de la regresión lineal múltiple

Para la parte de la regresión lineal múltiple, se aplicaron diferentes modelos para ver cómo algunas variables (como número de camas, reseñas o alojamientos) influyen sobre otras, como la tasa de aceptación del anfitrión o el total de listados.

Voy a explicar los diferentes modelos que realicé, algunos tienen el código comentado porque no se apreciaban bien en las gráficas (haciendo que se vieran muy dispersas y no se encontraran los puntos) y eso me hizo deducir que no tenían relación, por eso las comenté.

Modelo 1

Variables: accommodates, bedrooms, reviews_per_month = host_acceptance_rate

$R^2 = 0.027$

Con un R^2 de 0.027, el modelo explica solo el 2.7% de la variabilidad de la variable dependiente. Esto indica que no hay una relación significativa entre las variables elegidas y el comportamiento del anfitrión al aceptar solicitudes. Probablemente, la tasa de aceptación está más influenciada por factores personales.

Modelo 2

Variables: host_acceptance_rate, accommodates, bedrooms = host_total_listings_count

$R^2 = 0.011$

El resultado fue un R^2 de 0.011, lo cual significa que el modelo explica solo el 1.1% de la variabilidad de la variable dependiente. Es probable que el número de listados esté más relacionado con factores externos como si se trata de un anfitrión profesional, empresa o persona individual, algo que no se refleja en las variables usadas aquí.

Modelo 3: mejor

Variables: host_acceptance_rate, host_total_listings_count, bedrooms = accommodates

$R^2 = 0.360$

Este modelo predice la capacidad del alojamiento (número de personas que puede recibir) a partir del comportamiento del anfitrión y características del alojamiento. Con un R^2 de 0.36, el modelo explica aproximadamente el 36% de la variabilidad en la capacidad. El número de habitaciones (bedrooms) y la cantidad de propiedades que maneja un anfitrión pueden estar relacionados con la capacidad total que ofrecen. Es decir, un anfitrión con más habitaciones y más listados suele poder alojar a más personas.

Modelo 4: mejor

Variables: host_acceptance_rate, accommodates, reviews_per_month = bedrooms

$R^2 = 0.364$

Con un R^2 de 0.364, se explica aproximadamente el 36.4% de la variabilidad en el número de habitaciones. Es decir, hay una relación significativa, especialmente porque tiene sentido que alojamientos con más capacidad suelan tener más habitaciones. Además, una mayor frecuencia de reseñas podría estar relacionada con alojamientos más grandes o con más rotación de huéspedes, lo cual también puede estar vinculado al número de cuartos.

Modelo 5

Variables: host_acceptance_rate, host_total_listings_count, accommodates, bedrooms = price

$R^2 = 0.0157$

El resultado, con un R^2 del 1.57%, muestra que no hay una relación significativa entre estas variables y el precio. Es decir, el precio no depende tanto de cuántas personas caben ni del número de habitaciones o alojamientos del anfitrión, sino probablemente de factores como la ubicación, temporada, tipo de alojamiento o estrategia del anfitrión.

Modelo 6

Variables: host_acceptance_rate, host_total_listings_count, accommodates, bedrooms = reviews_per_month

$R^2 = 0.037$

Con un R^2 de 3.7%, la capacidad del modelo es muy baja. Esto indica que estas variables no explican bien el volumen de reseñas mensuales, ya que probablemente este dato depende más de la experiencia del huésped, ubicación, atención o reputación en línea que de los elementos considerados aquí.