Tecnológico de Monterrey

Actividad 1

Raymundo Díaz Tijera A01664497

28 sept 2025

Plataformas de analítica de negocios para organizaciones

Profesor: Maria Luisa Gomez Barrios

Co-Titular: Alfredo García Suárez

Se trabajará con la base de datos de Berlin, Alemania Airbnb, la cual cuenta con:

- 77 columnas
- 14,187 filas

Se realizará un análisis mediante regresión lineal y múltiple para identificar diversos hallazgos y comportamientos.

Punto 4. Correlación entre los 4 tipos de la columna 'room_type':

- Entire home/apt

Dependiente	Independiente	Correlación
host_acceptance_rate	host_response_rate	0.36
host_acceptance_rate	price	0.07
host_acceptance_rate	number_of_reviews	0.19
review_scores_rating	calcluated_host_listings_count	0.16
availability_365	number_of_reviews	0.13
reviews_per_month	review_scores_communication	0.39

- Private room

Dependiente	Independiente	Correlación
host_acceptance_rate	host_response_rate	0.29
host_acceptance_rate	price	0.16
host_acceptance_rate	number_of_reviews	0.13
review_scores_rating	calcluated_host_listings_count	0.28
availability_365	number_of_reviews	0.07
reviews_per_month	review_scores_communication	0.35

- Hotel room

Dependiente	Independiente	Correlación
host_acceptance_rate	host_response_rate	0.19
host_acceptance_rate	price	0.03
host_acceptance_rate	number_of_reviews	0.27
review_scores_rating	calcluated_host_listings_count	0.28
availability_365	number_of_reviews	0.05
reviews_per_month	review_scores_communication	0.29

- Shared room

Dependiente	Independiente	Correlación
host_acceptance_rate	host_response_rate	0.43
host_acceptance_rate	price	0.08
host_acceptance_rate	number_of_reviews	0.19
review_scores_rating	calcluated_host_listings_count	0.12
availability_365	number_of_reviews	0.07
reviews_per_month	review_scores_communication	0.45

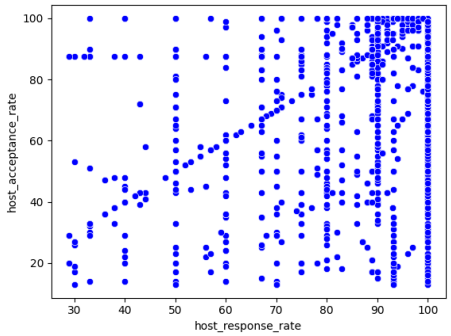
El par de variables con más variabilidad en su correlación a través de las 4 categorías de room_type, fue la **variable dependiente (y) 'host_acceptance_rate'**, con la **variable independiente (x) 'host_response_rate'**.

- Su rango (máx - mín) va de la correlación entre Shared room (0.43) y Hotel room (0.19) = 0.24.

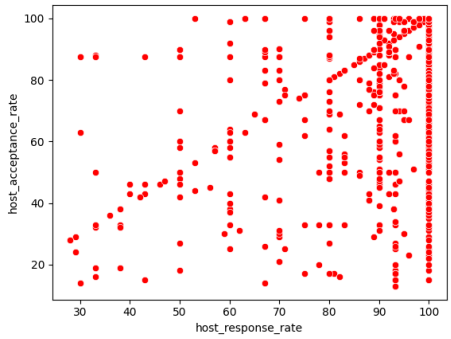
El segundo par con mayor variabilidad fue la variable dependiente 'reviews_per_month' con la variable independiente 'review_scores_communication', con un rango entre las 4 categorías de Shared room (0.45) y Hotel room (0.29) = 0.16.

Para el análisis se decide optar por realizar 4 gráficos de dispersión con ‘host_acceptance_rate’ y ‘host_response_rate’, para las 4 categorías previamente definidas.

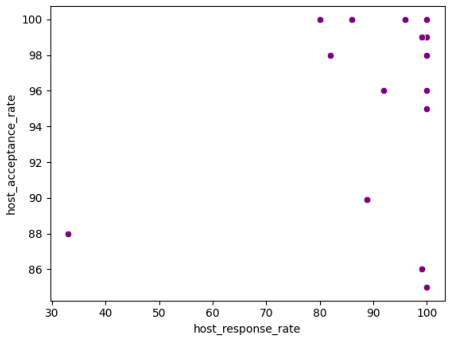
Entire home/apt (0.36)



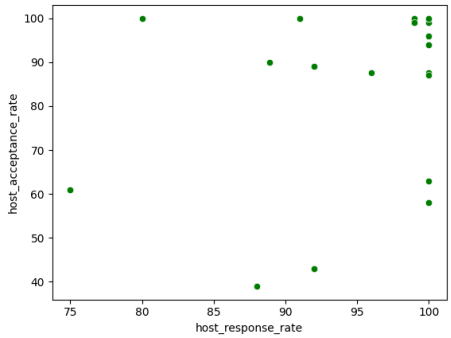
Private room (0.29)



Hotel room (0.19)



Shared room (0.43)



Hallazgos:

- Las relaciones entre host_acceptance_rate (Y) y host_response_rate (x) no son uniformes para las 4 categorías de room_type, teniendo una gran variabilidad de datos que puede verse reflejado en las correlaciones obtenidas, al ser todas menores a 0.50, siendo Shared room la correlación más alta con un 0.43,
- Lo que puede implicar que los hosts tienden a contestar más pronto, suelen aceptar más reservas en Shared room, que puede ser el tipo de categoría en donde impacte más el comportamiento de los hosts, con una mayor probabilidad de respuesta para los huéspedes.

- Para hotel room, con su correlación de 0.19, siendo la menor, podría implicar que este tipo de habitación tiene procesos estandarizados de aceptación, sin tener tanta influencia por parte de la aceptación de los hosts.
- Entire home y Private room son las categorías con mayor número de registros (9580 y 4397 respectivamente), con una correlación media positiva, que puede indicar que la tasa de respuesta por parte del host, influye de una manera no significativa para la aceptación de la misma, que podría ser que hay diversos hosts que aceptan o no, sin importar si se tardan o no, y otros casos donde responden rápido pero no aceptan la reserva.

Punto 5. 10 variables con mayor correlación para cada tipo de alojamiento en room_type

- **Entire home/apt**

1. host_listings_count host_total_listings_count 0.996542
2. review_scores_rating review_scores_accuracy 0.995612
3. review_scores_checkin review_scores_communication 0.993950
4. review_scores_accuracy review_scores_communication 0.993736
5. review_scores_rating review_scores_checkin 0.993364
6. review_scores_accuracy review_scores_checkin 0.993008
7. review_scores_rating review_scores_cleanliness 0.992799
8. review_scores_accuracy review_scores_cleanliness 0.991777
9. review_scores_communication review_scores_value 0.991708
10. review_scores_accuracy review_scores_location 0.991287

- **Private room**

1. host_listings_count host_total_listings_count 0.999107
2. minimum_minimum_nights minimum_nights_avg_ntm 0.994729
3. minimum_nights minimum_nights_avg_ntm 0.994329
4. review_scores_checkin review_scores_communication 0.993085
5. review_scores_rating review_scores_accuracy 0.993082
6. review_scores_accuracy review_scores_communication 0.991926
7. review_scores_rating review_scores_communication 0.991325
8. review_scores_accuracy review_scores_value 0.990922

9. review_scores_rating review_scores_value 0.990714
10. maximum_minimum_nights minimum_nights_avg_ntm 0.990655

- **Hotel room**

1. minimum_nights maximum_minimum_nights 0.996268
2. maximum_minimum_nights minimum_nights_avg_ntm 0.993454
3. minimum_nights minimum_nights_avg_ntm 0.993384
4. review_scores_rating review_scores_cleanliness 0.990382
5. review_scores_accuracy review_scores_value 0.987577
6. review_scores_rating review_scores_value 0.987335
7. review_scores_cleanliness review_scores_value 0.982275
8. review_scores_accuracy review_scores_location 0.982244
9. review_scores_location review_scores_value 0.981428
10. review_scores_accuracy review_scores_checkin 0.98132

- **Shared room**

1. minimum_nights minimum_nights_avg_ntm 0.999956
2. maximum_minimum_nights minimum_nights_avg_ntm 0.999603
3. minimum_nights maximum_minimum_nights 0.999453
4. host_listings_count calculated_host_listings_count 0.996018
5. review_scores_accuracy review_scores_checkin 0.995308
6. review_scores_rating review_scores_accuracy 0.994795
7. review_scores_checkin review_scores_location 0.994243
8. review_scores_accuracy review_scores_value 0.993384
9. review_scores_rating review_scores_communication 0.993249
10. review_scores_checkin review_scores_value 0.991293

Hallazgos:

- Hay correlaciones muy positivas, altas y fuertes, lo que indica que se debe tener cuidado para poder crear nuestros modelos de regresión múltiple, pues la multicolinearidad será un factor a tener en cuenta, y evitar sesgos en los modelos, en donde técnicas como Ridge/Lasso regression podrían ser de utilidad.

- Las correlaciones más altas se encuentran en reviews y minimum-maximum nights, lo que podría indicar que los huéspedes califican muy parecido o tienen comportamientos muy similares en las columnas con estos nombres, por lo que para los modelos utilizando estas variables, solo una variable será lo más recomendable a utilizar, ya que son bastante representativas.

Punto 6. Modelos de regresión múltiple, en base a las variables con mayor correlación:

Se excluirán las variables con 0.98 de correlación con la dependiente, para evitar multicolinearidad, evitar sesgos y obtener una representatividad explicativa real (no inflada).

1. y = review_scores_rating

   Se excluyen las variables review_scores_accuracy, review_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location, review_scores_value, todas con una correlación arriba de 0.99.

   X's = number_of_reviews_ltm (0.32), estimated_occupancy_l365d (0.43), estimated_revenue_l365d (0.35), y reviews_per_month (0.38).

   R^2: 0.201

   R: 0.448

2. y = host_acceptance_rate

   X's = host_response_rate (0.34), reviews_per_month (0.21), y estimated_revenue_l365d (0.19)

   R^2: 0.13

   R: 0.36

3. host_is_superhost fue sustituida por availability_30

   y = availability_30

   X's = availability_eoy (0.78), calculated_host_listings_count (0.20) y maximum_nights (-0.20)

   R^2: 0.61

   R: 0.78

4. y = host_total_listings_count

   X's = calculated_host_listings_count (0.46), minimum_nights_avg_ntm (0.14), y review_scores_communication (-0.31)

   R^2: 0.27

   R: 0.52

5. y = accommodates

   X's = beds (0.70), estimated_revenue_l365d (0.30), price (0.29), y bathrooms (0.26)

   R^2: 0.52

   R: 0.72

6. y = bedrooms

   X's = accommodates (0.55), beds (0.50), bathrooms (0.32), y price (0.25)

   R^2: 0.37

   R: 0.61

7. y = Price

   X's = accommodates (0.29), bedrooms (0.25), estimated_revenue_l365d (0.23), y bathrooms (0.20)

   R^2: 0.13

   R: 0.36

8. y = review_scores_value

   Se excluyen las variables review_scores_accuracy, review_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location, review_scores_value, todas con una correlación arriba de 0.98, para evitar tener un modelo con multicolinearidad.

   X's = estimated_occupancy_l365d (0.42), reviews_per_month (0.38), y number_of_reviews (0.31).

   R^2: 0.19

   R: 0.43

9. y = bathrooms

   X's = bedrooms (0.33), accommodates (0.26), y price (0.20)

R^2: 0.13

R: 0.36

10. y = reviews_per_month

    X's = number_of_reviews_ltm (0.81), estimated_occupancy_l365d (0.76), review_scores_rating (0.38), y host_response_rate (0.29)

    R^2: 0.69

    R: 0.83

Hallazgos:

Los modelos se pueden clasificar en tres categorías:

1. Moderada - Fuerte representatividad

   reviews_per_month fue el mejor modelo hecho entre todas las variables, con una r cuadrada de 0.69, lo que nos indica que las variables independientes explican el 69% de la variabilidad para el número de reseñas por mes, mayormente influenciado por el historial de reseñas y ocupación estimada en los 365 días.

   availability_30 fue el segundo mejor modelo, con una r cuadrada de 0.61, con una disponibilidad futura moderadamente influenciada por la disponibilidad anual, penalizada por el número máximo de noches permitido.

   accommodates fue el tercer mejor modelo con una r cuadrada de 0.52, por lo que beds, bathrooms, revenue y price determinan en un 52% la variabilidad de la disponibilidad para la capacidad alojamiento.

   Los tres modelos tienen una R mayor a 0.70, lo que indica un buen coeficiente de correlación moderada-buena.

2. Moderada representatividad

   bedrooms, con una r cuadrada de 0.37, es consistente, que entre más camas y capacidad suelen reflejar más dormitorios, pero solo para el 37% de las ocasiones.

   host_total_listings_count, con una r cuadrada de 0.27, que se podría interpretarse como una dependencia baja-moderada de noches mínimas y reviews sobre el número total de

propiedades por host, donde otras variables cualitativas podrían tener un poder explicativo mayor.

Ambas tienen un coeficiente de correlación mayor a 0.50, lo que indica todavía una correlación moderada moderada,

3. Baja representatividad

   review_scores_rating, con una r cuadrada de 0.201, al excluir las variables con 0.99 de correlación, se trabajó con variables más operativas como estimated revenue y estimated occupancy, detectando que solo pueden interpretar el 20% de la variabilidad para esta variable de satisfacción para huésped.

   review_scores_value, con una r cuadrada de 0.19, al excluir también las variables con 0.99, se trabajó con variables operativas, pero de igual manera no son buenas variables predictoras.

   price, con una r cuadrada de 0.13, las variables de capacidad, dormitorios y baños tienen una representatividad débil para el precio, ya que puede ser que dependan más sobre estrategias de promoción y ubicación.

   host_acceptance_rate, con una r cuadrada de 0.13, en general, la tasa de aceptación depende más de decisión individuales del host, no de la tasa de respuesta, métricas cuantitativas de alguna promoción o ingresos estimados.

   bathrooms, con una r cuadrada de 0.13, el número de baños probablemente depende más de la ubicación o ingresos estimados, que sobre dormitorios, su capacidad y el mismo precio.

   Todos los modelos tienen un coeficiente de correlación menor a 0.40, lo que indica una correlación débil - moderada.

En conclusión

Los modelos con mejor representatividad son aquellos donde la relación entre variables es más cuantificable, mientras que los modelos con baja representatividad son los relacionados con percepciones o decisiones del host que dependen de factores intangibles y externos a variables cuantitativas, donde un análisis de variables categóricas como el que hicimos el bloque pasado,

es de suma importancia para entender aún más el comportamiento de los huéspedes en Berlín Alemania.