Instituto Tecnológico y de Estudios Superiores de Monterrey

Actividad 1

Reporte

José Francisco Vera Jiménez

1. Carga y Exploración Inicial de Datos

Se utilizó el dataset correspondiente a la ciudad de Estocolmo (stocolmo_inicial.csv), obtenido desde *Inside Airbnb*. La base fue importada en un entorno de Google Colab utilizando la librería **pandas** para su manipulación.

```
import pandas as pd

df = pd.read_csv('/content/stocolmo_inicial.csv')

display(df.head())
```

Inicialmente, se realizó una revisión general para identificar el número de valores nulos presentes en cada columna.

El resultado mostró que variables como neighbourhood_overview, host_about, calendar_updated, license y neighbourhood_group_cleansed presentaban un número considerable de datos faltantes.

Entre las variables numéricas, bathrooms, beds, price y estimated_revenue_l365d también mostraban valores nulos relevantes.

2. Imputación de Valores Faltantes

Con el propósito de conservar la integridad de la base y evitar eliminar registros, se aplicaron distintas estrategias de imputación:

- **Variables de texto:** se reemplazaron los valores nulos con la cadena "No Information".
- **Variables numéricas:** se reemplazaron los valores faltantes con la **mediana** de cada columna, una técnica adecuada para mitigar la influencia de valores extremos.

El bloque de código responsable de esta etapa fue:

```
cols_texto_elegidas =
["name","description","neighborhood_overview","host_name","host_location",

"host_about","host_response_time","host_response_rate","host_acceptance_rate","host_since",

"host_thumbnail_url","host_picture_url","has_availability","host_neighbourhood","neighbourhood",
```

"neighbourhood_cleansed","property_type","room_type","bathrooms_text","calendar_updated",

"first_review","last_review","license","host_is_superhost","host_has_profile_pic",

"host_identity_verified","neighbourhood_group_cleansed","host_verifications"]

df[cols_texto_elegidas] = df[cols_texto_elegidas].fillna("No Information")

Posteriormente, se convirtieron las columnas de fechas (host_since, first_review, last_review) al formato datetime para su correcta manipulación:

cols_fechas = ["host_since", "first_review", "last_review"]

for col in cols_fechas:

  df[col] = pd.to_datetime(df[col], errors="coerce")

El análisis de rangos mostró que las fechas de alta de anfitriones (host_since) abarcan desde noviembre de 2008 hasta junio de 2025, mientras que las reseñas (first_review, last_review) cubren el periodo de 2011 a 2025.

3. Limpieza de Datos Monetarios y Numéricos

Las variables **price** y **estimated_revenue_l365d** se encontraban en formato texto con símbolos monetarios y comas. Se procedió a limpiar dichas columnas y convertirlas a tipo numérico (float):

df["price"] = df["price"].astype(str).str.replace("[$,]", "", regex=True).astype(float)

df["estimated_revenue_l365d"] = df["estimated_revenue_l365d"].astype(str).str.replace("[$,]", "", regex=True).astype(float)

Se imputaron los valores faltantes de variables cuantitativas mediante la mediana, entre ellas: bathrooms, bedrooms, beds, price, estimated_revenue_l365d y los distintos mínimos y máximos de noches de estancia.

Tras esta imputación, las únicas columnas que conservaron nulos fueron las correspondientes a fechas (host_since, first_review, last_review), debido a la ausencia de registros originales.

4. Limpieza de Outliers

Para garantizar la homogeneidad de los datos, se eliminaron los valores atípicos mediante el método de desviación estándar (3σ).

Se implementó además una interfaz interactiva en Colab para visualizar los *boxplots* y detectar visualmente los outliers antes de imputar

Después del proceso, el dataset pasó de 5,315 registros originales a 4,161 registros, lo que indica una eliminación aproximada del 21.7 % de observaciones consideradas atípicas.

5. Conversión de Porcentajes y Preparación Final

Se estandarizaron las columnas con valores porcentuales (host_acceptance_rate, host_response_rate) eliminando el símbolo % y convirtiéndolas a tipo numérico:

```
cols_porcentaje = ["host_acceptance_rate", "host_response_rate"]

for col in cols_porcentaje:
  df_clean[col] = (
    df_clean[col]
    .astype(str)
    .str.replace("%", "", regex=True)
    .replace("No Information", np.nan)
    .astype(float)
  )
```

Se verificó su correcta conversión con resultados como:

host_acceptance_rate float64 [83., 35., 0., 100., 13., 43., 69., 73., nan, 40.]

host_response_rate float64 [80., 0., 100., nan, 57., 75., 56., 83., 33., 60.]

Análisis de Regresión Lineal Simple

Una vez que la base de datos fue depurada y estandarizada, se procedió al desarrollo del análisis de Regresión Lineal Simple, cuyo objetivo fue evaluar la relación entre variables clave del alojamiento y su impacto en diferentes tipos de habitación.

Para ello, se utilizaron las bibliotecas *Seaborn*, *Matplotlib* y *Scikit-Learn*, que permiten estimar el modelo lineal, calcular el coeficiente de determinación ($\mathbf{R^2}$) y generar los gráficos de dispersión con la línea de ajuste.

El procedimiento consistió en iterar por cada tipo de habitación (Entire home/apt, Private room, Shared room y Hotel room) y evaluar seis combinaciones de variables en pares dependiente–independiente.

En particular, el tipo de habitación *Shared room* presentó una cantidad muy limitada de registros válidos tras la limpieza de datos y la eliminación de outliers. Esta escasa muestra impidió realizar regresiones lineales confiables, ya que el número de observaciones fue menor a cinco. Por motivos de rigor metodológico y para evitar interpretaciones erróneas, se optó por no incluir este tipo de alojamiento en los resultados de correlación y regresión, dejando constancia de su baja representatividad dentro del conjunto de datos analizado.

Estos pares fueron definidos con base en su relevancia operativa dentro de la plataforma Airbnb y su potencial para explicar comportamientos del mercado:

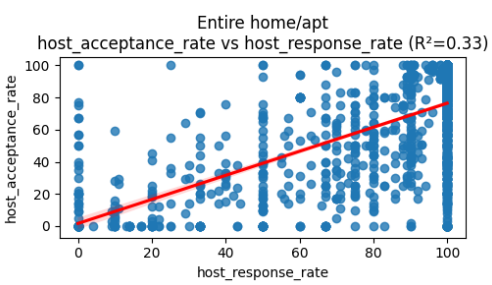
1. **host_acceptance_rate vs host_response_rate**

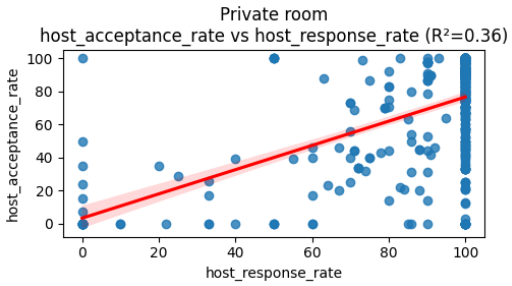
   Esta regresión analizó la relación entre la tasa de aceptación y la tasa de respuesta de los anfitriones.

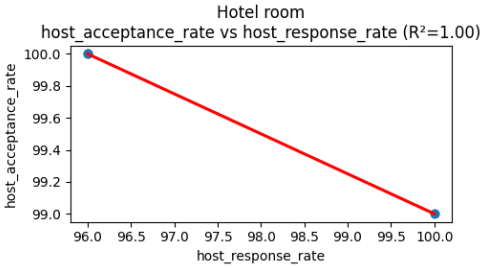
   En Entire home/apt (R^2 = 0.33) y Private room (R^2 = 0.36) se observó una correlación positiva moderada, lo que indica que los anfitriones que responden con mayor frecuencia también suelen aceptar más solicitudes.

   En Hotel room, el valor R^2 = 1.00 no es representativo, ya que el número de observaciones fue muy bajo.

   Los resultados sugieren que la actividad y disposición del anfitrión influyen positivamente en sus tasas de aceptación, excepto en categorías con pocos datos como los hoteles.





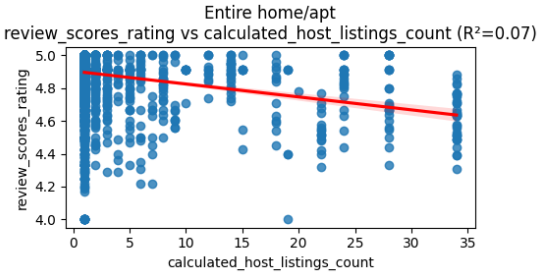


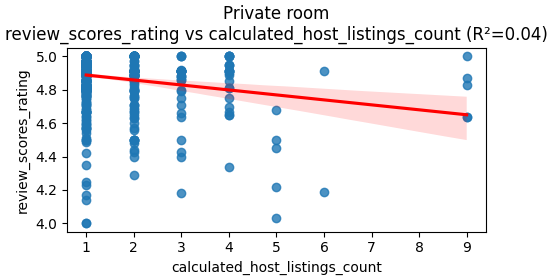
2. **review_scores_rating vs calculated_host_listings_count**
   Esta regresión evaluó si el número de propiedades administradas por un anfitrión influye en sus calificaciones.

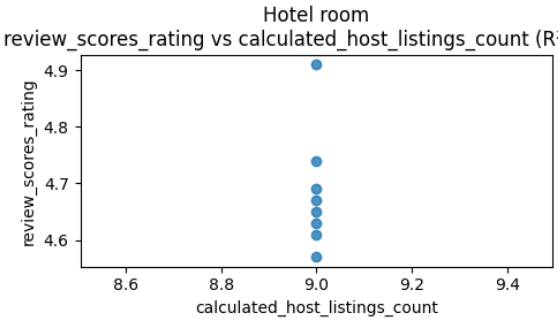
   Los resultados muestran una relación débil y ligeramente negativa tanto en Entire home/apt (R^2 = 0.07) como en Private room (R^2 = 0.04), indicando que manejar más alojamientos no garantiza mejores evaluaciones.

   En Hotel room (R^2 = 0.00) no se detectó relación, debido a la falta de variabilidad en las calificaciones.

   En conjunto, se concluye que la cantidad de propiedades no determina la satisfacción del huésped, la cual depende más de factores cualitativos como la atención o la limpieza.



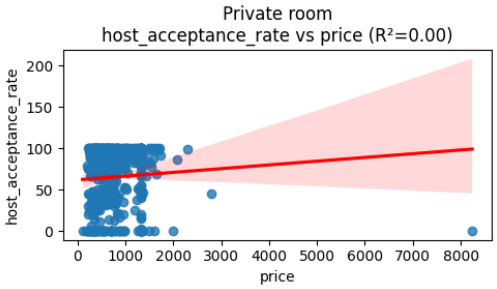


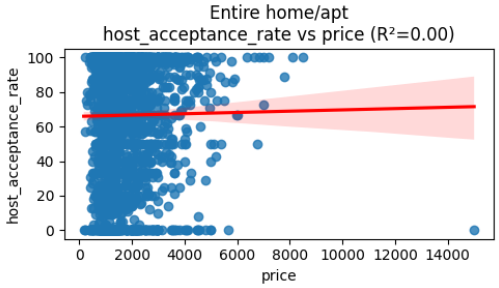


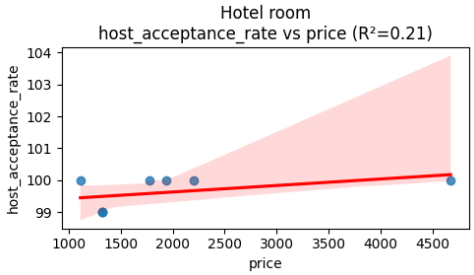
3. **host_acceptance_rate vs price**
   El análisis de la relación entre la tasa de aceptación y el precio del alojamiento muestra una dispersión amplia y sin una tendencia clara en los tipos Entire home/apt y Private room. En ambos casos (R^2 = 0.00), los puntos se concentran en la parte inferior del eje de precios, indicando que la mayoría de los alojamientos mantienen precios bajos o moderados sin que esto afecte la disposición del anfitrión a aceptar reservas.

Visualmente, las nubes de puntos son densas y sin pendiente definida, lo que confirma la ausencia de relación lineal entre ambas variables.

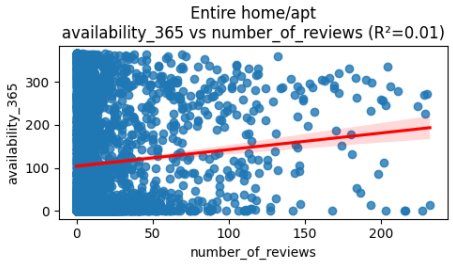
En Hotel room, aunque el coeficiente de determinación ($R^2 = 0.21$) es ligeramente superior, las observaciones son escasas. La línea de tendencia presenta una leve pendiente positiva, pero el patrón sigue siendo disperso, por lo que no se puede concluir que un precio mayor esté asociado con una tasa de aceptación más alta.

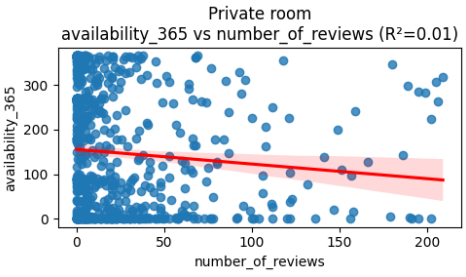


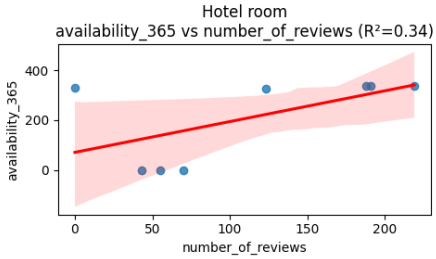




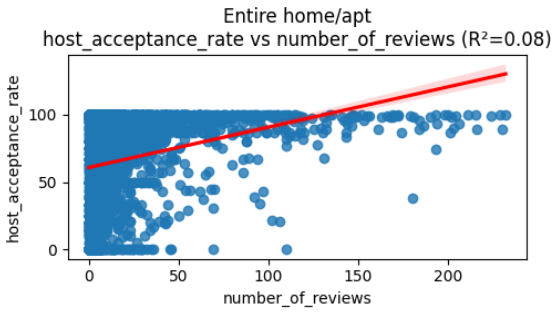
4. **availability_365 vs number_of_reviews**
   Determina si las propiedades más disponibles durante el año tienden a recibir un mayor número de reseñas.

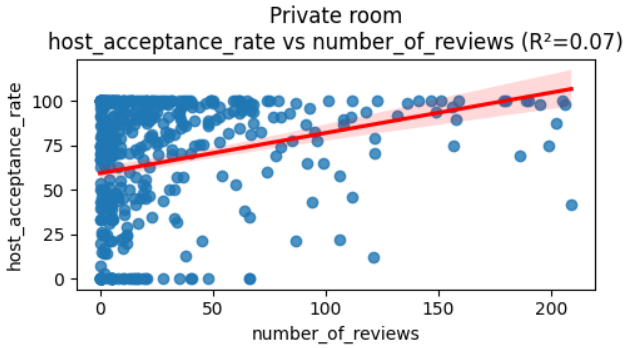


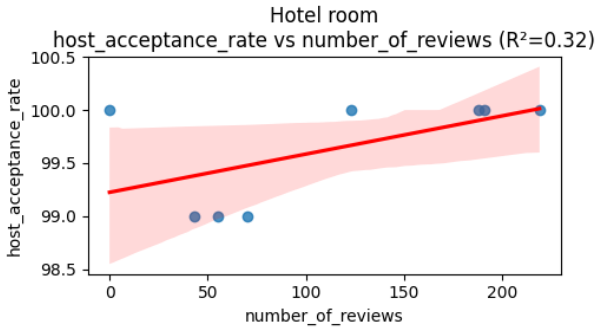




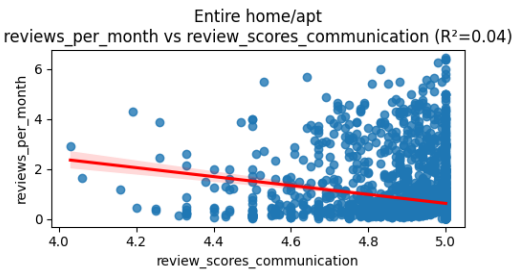
5. **host_acceptance_rate vs number_of_reviews**
   Busca identificar si los anfitriones con alta tasa de aceptación acumulan más comentarios.

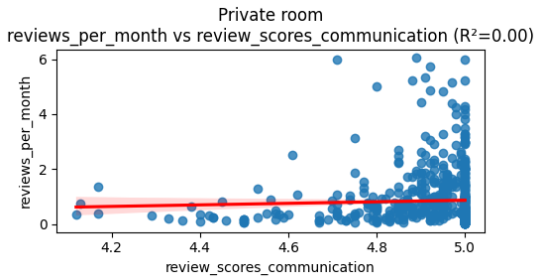


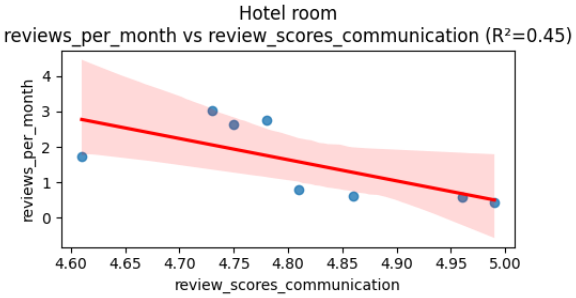




6. **reviews_per_month vs review_scores_communication**
   Explora la relación entre la frecuencia de reseñas mensuales y la calidad de la comunicación reportada por los huéspedes.







Para cada combinación se calculó el coeficiente de determinación (R^2), indicador que mide qué tan bien la variable independiente explica la variabilidad de la dependiente. Además, se generaron los gráficos de dispersión correspondientes con la línea de tendencia en color rojo, lo que permitió observar visualmente la dirección y fuerza de la relación.

9. Análisis de Correlaciones mediante Mapas de Calor

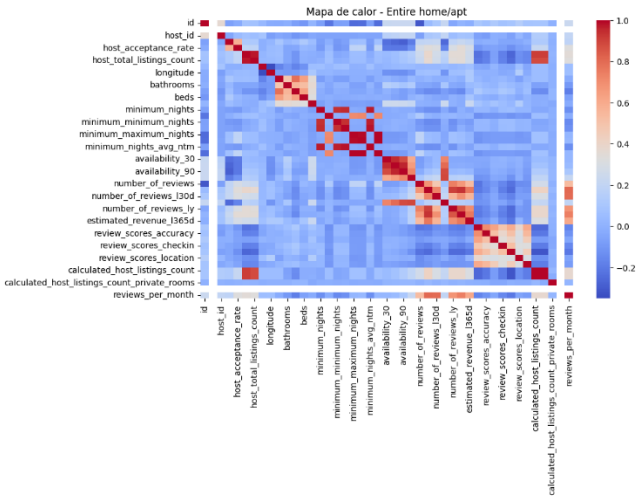
Con el fin de identificar las relaciones más relevantes entre las variables numéricas del conjunto de datos, se elaboraron mapas de calor (heatmaps) segmentados por tipo de habitación.
Esta visualización permite observar de manera global cómo se comportan las correlaciones, resaltando las variables que presentan asociaciones más fuertes, tanto positivas como negativas.

El análisis se realizó para los cuatro tipos de alojamiento definidos en la base de datos: *Entire home/apt*, *Private room*, *Shared room* y *Hotel room*. Para cada categoría se generó una matriz de correlación utilizando las variables numéricas depuradas durante el preprocesamiento.

Adicionalmente, se calcularon las 10 correlaciones con mayor valor absoluto para cada tipo de habitación, lo que permitió identificar los pares de variables con relaciones más destacadas dentro de cada segmento. En los casos con menos de 50 registros válidos principalmente *Shared room* se omitió la generación del mapa, debido a la falta de datos suficientes para obtener correlaciones estadísticamente confiables.

Este análisis es fundamental para comprender la estructura interna del dataset y orientar la construcción de los modelos de regresión múltiple en etapas posteriores.

Mapa de Calor: Entire home/apt



El mapa de calor correspondiente al tipo de alojamiento *Entire home/apt* muestra una matriz de correlaciones con patrones bien definidos entre diversas variables numéricas. Se observa una alta correlación positiva entre las variables asociadas a la disponibilidad como availability_30, availability_90 y availability_365, lo que indica que los niveles de ocupación a corto, mediano y largo plazo tienden a comportarse de manera similar.

Asimismo, existe una relación significativa entre las variables de reseñas (number_of_reviews, number_of_reviews_l30d, number_of_reviews_ly y reviews_per_month), lo cual refleja consistencia en la actividad de los huéspedes a lo largo del tiempo.

Por otro lado, variables como price, review_scores_rating y host_acceptance_rate presentan correlaciones bajas con la mayoría de los indicadores, lo que sugiere que en los alojamientos completos el precio o la reputación del anfitrión no guardan una relación directa con la frecuencia de reservas ni con la disponibilidad.

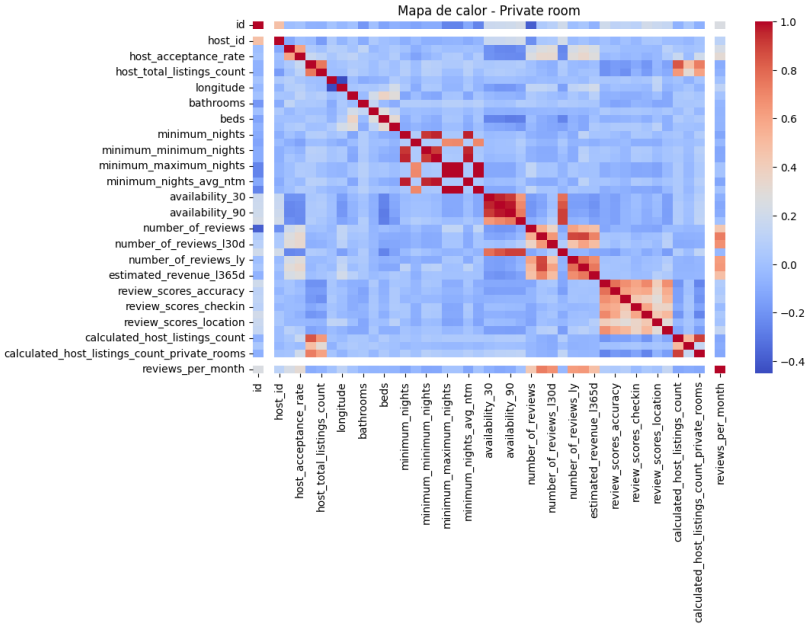
Este mapa revela una estructura interna coherente y homogénea, donde las variables operativas de disponibilidad y reseñas mantienen los vínculos más fuertes, mientras que los aspectos de valoración y precio muestran una independencia marcada.

Mapa de Calor: Private room

El mapa de calor correspondiente a *Private room* muestra una red de correlaciones moderadas entre variables de disponibilidad y actividad de huéspedes. Se aprecia una fuerte asociación entre availability_30, availability_90 y availability_365, lo que confirma que los periodos de ocupación a corto, mediano y largo plazo tienden a variar en conjunto.

Asimismo, se observan correlaciones positivas entre number_of_reviews, reviews_per_month y estimated_revenue_l365d, lo que indica que las habitaciones privadas con mayor rotación de huéspedes registran también más reseñas y mayores ingresos estimados.

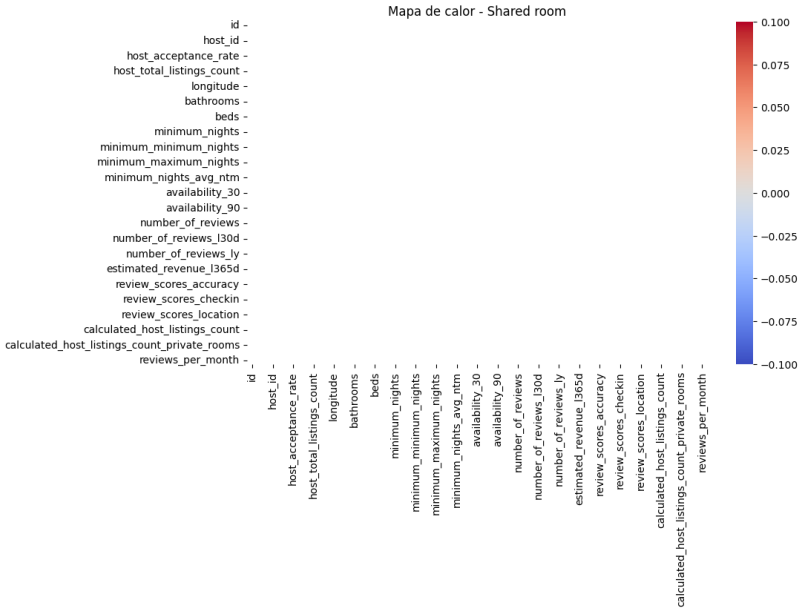
Por el contrario, variables como price, review_scores_rating y host_acceptance_rate presentan baja correlación con el resto, evidenciando que los precios y las calificaciones no dependen directamente de la disponibilidad o del número de reseñas.



Mapa de Calor: Shared room

En el caso de *Shared room*, el mapa de calor no muestra patrones de correlación definidos, reflejando una ausencia de relaciones significativas entre las variables numéricas analizadas. Esto se debe principalmente al bajo número de registros disponibles para este tipo de alojamiento después del proceso de limpieza y eliminación de outliers.

La escasez de datos impide obtener correlaciones estables o representativas, lo que genera una matriz prácticamente vacía. Este resultado evidencia que la categoría de habitaciones compartidas tiene una presencia muy limitada dentro del conjunto de datos, por lo que sus valores no permiten realizar inferencias estadísticas confiables.



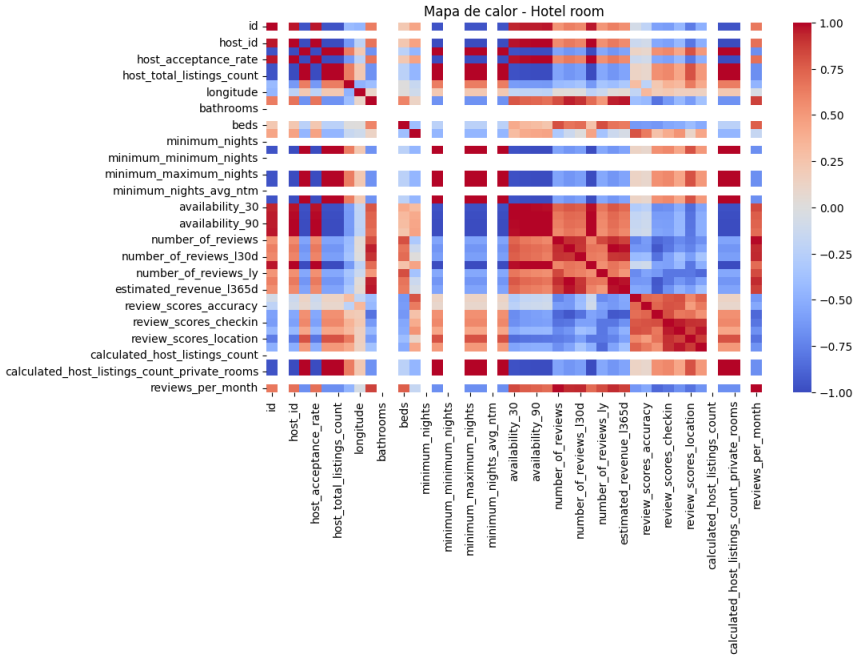
Mapa de Calor: Hotel room

El mapa de calor correspondiente al tipo de alojamiento *Hotel room* presenta un patrón de correlaciones más fragmentado, con zonas de alta asociación entre grupos de variables específicas.

Se observan correlaciones positivas fuertes entre las variables de reseñas (number_of_reviews, number_of_reviews_l30d, number_of_reviews_ly, reviews_per_month) y los ingresos estimados anuales (estimated_revenue_l365d), lo que indica que los hoteles con mayor rotación de huéspedes tienden a generar más ingresos y acumular más comentarios.

Asimismo, las variables de disponibilidad (availability_30, availability_90, availability_365) también muestran vínculos consistentes entre sí, aunque con cierta variabilidad, lo que podría reflejar diferencias estacionales o de gestión en las reservas hoteleras.

Por otro lado, algunas correlaciones negativas aisladas entre minimum_nights y las variables de reseñas sugieren que los hoteles con estancias mínimas más largas tienden a recibir menos comentarios, posiblemente por tener menor rotación de huéspedes.



Modelos de Regresión Lineal Múltiple

Durante esta etapa del proyecto se realizó la creación de distintos modelos de regresión lineal múltiple para analizar las relaciones entre diversas variables cuantitativas del conjunto de datos de Airbnb en Estocolmo. El objetivo fue evaluar qué tan bien las variables independientes podían explicar la variabilidad de cada variable dependiente seleccionada.

Primero se preparó la estructura donde se almacenarían los resultados de cada modelo. Posteriormente, se configuró un ciclo que recorrió todas las variables seleccionadas y, para cada una de ellas, se construyó un modelo de regresión lineal múltiple utilizando las demás variables numéricas disponibles como predictoras.

Antes de iniciar el modelado, fue necesario transformar la variable **“host_is_superhost”**, ya que originalmente se encontraba en formato de texto (“t” y “f”). Para que el modelo pudiera procesarla correctamente, se convirtió a valores numéricos binarios, asignando 1 para los anfitriones superhost y 0 para los que no lo eran. De esta manera, se garantizó que todas las variables utilizadas fueran numéricas y compatibles con el algoritmo.

Posteriormente, se seleccionaron únicamente las columnas de tipo numérico para el análisis, eliminando aquellas observaciones que contaban con valores nulos. Este paso permitió trabajar con datos consistentes y evitar errores durante el proceso de entrenamiento de los modelos.

Una vez preparada la base de datos, se entrenó un modelo de regresión lineal múltiple para cada una de las variables cuantitativas especificadas. En cada caso, se evaluó el coeficiente de determinación (R^2), el cual indica qué proporción de la variabilidad de la variable dependiente puede ser explicada por las variables independientes consideradas.

Los resultados obtenidos fueron los siguientes:

```
✅ Modelo para 'review_scores_rating' completado | R²=0.032
✅ Modelo para 'host_acceptance_rate' completado | R²=0.004
✅ Modelo para 'host_is_superhost' completado | R²=0.013
✅ Modelo para 'host_total_listings_count' completado | R²=0.005
✅ Modelo para 'accommodates' completado | R²=0.003
✅ Modelo para 'bedrooms' completado | R²=0.014
✅ Modelo para 'price' completado | R²=0.005
✅ Modelo para 'review_scores_value' completado | R²=0.016
✅ Modelo para 'bathrooms' completado | R²=0.019
✅ Modelo para 'reviews_per_month' completado | R²=0.055
```

Durante este procedimiento se construyeron distintos modelos de regresión lineal múltiple, uno para cada variable cuantitativa del conjunto de datos. El propósito fue determinar qué tan bien las demás variables numéricas del dataset podían explicar el comportamiento de cada variable objetivo, mediante el cálculo del coeficiente de determinación (R^2).

Después de ejecutar los modelos y organizar los resultados en orden descendente, se obtuvieron los siguientes valores de R^2:

	Variable dependiente	R²
9	reviews_per_month	0.055346
0	review_scores_rating	0.032457
8	bathrooms	0.019296
7	review_scores_value	0.016071
5	bedrooms	0.014113
2	host_is_superhost	0.013315
6	price	0.005086
3	host_total_listings_count	0.004659
1	host_acceptance_rate	0.003558
4	accommodates	0.002558

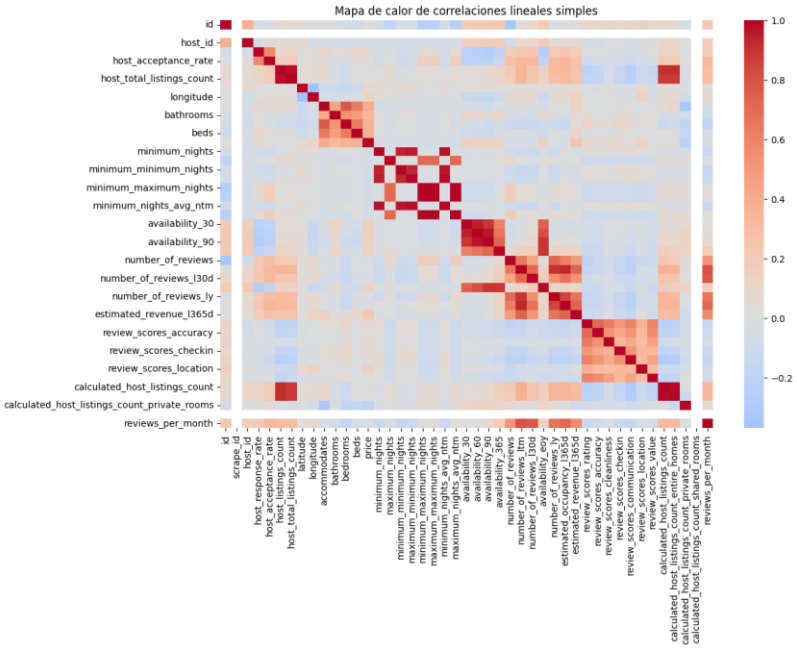
Los resultados muestran que, en todos los casos, los valores de R^2 son relativamente bajos, lo que significa que las variables independientes disponibles no explican de manera significativa la variabilidad de las variables dependientes analizadas. En términos prácticos, el modelo lineal no logra capturar la complejidad del fenómeno descrito por los datos.

El modelo con mejor desempeño fue el de reviews_per_month, con un R^2 de 0.0553, lo que indica que solo el 5.5 % de la variabilidad en la cantidad de reseñas mensuales puede ser explicada por las demás variables del conjunto.
En contraste, el modelo con menor desempeño fue accommodates, con un R^2 de 0.0025, lo que sugiere una relación prácticamente nula con el resto de las variables.

El mapa de calor de correlaciones muestra las relaciones lineales simples entre las variables numéricas del conjunto de datos. Se observan correlaciones fuertes entre las variables relacionadas con las noches mínimas y máximas de estancia, así como entre los conteos de listados de los anfitriones, lo que indica colinealidad.

También se aprecian correlaciones moderadas entre variables como price, accommodates, bedrooms y bathrooms, lo cual es lógico, ya que los precios suelen aumentar con la capacidad y tamaño del alojamiento. En cambio, las variables de puntuaciones (review_scores_*) muestran relaciones mucho más débiles.

Al comparar estos resultados con los modelos de regresión lineal múltiple, se confirma que los valores de R^2 son bajos, lo que significa que, aunque existen relaciones simples entre algunas variables, al analizarlas de manera conjunta el modelo no logra explicar gran parte de la variabilidad. Esto se debe principalmente a la colinealidad y a la influencia de factores externos que no están representados en las variables numéricas disponibles.



Al analizar el mapa de calor y la tabla de correlaciones, se nota que las variables más relacionadas son las que tienen que ver con el tamaño del alojamiento, como accommodates, bedrooms y bathrooms, que muestran una relación positiva entre sí. También se ve una relación alta entre los conteos de listados de los anfitriones, lo que indica que esas variables repiten información similar.

Las variables de calificación (como *review_scores_rating* y *review_scores_value*) también se relacionan entre sí, lo que tiene sentido porque si un alojamiento recibe buenas reseñas en un aspecto, normalmente las obtiene en los demás.

Por otro lado, variables como price, host_acceptance_rate y reviews_per_month muestran correlaciones más bajas con el resto. Esto explica por qué los modelos de regresión lineal múltiple tuvieron valores de R^2 muy bajos, ya que las relaciones entre las variables no son tan fuertes o no siguen un patrón lineal claro.