Reporte de Actividad 2.1 (Regresión Lineal)

Preprocesamiento de los datos

Realice una limpieza del conjunto de datos original, incluyendo:

- Conversión de campos como host_acceptance_rate, host_response_rate y price a formato numérico.
- Imputación de valores nulos con la **media** en variables numéricas y con la **moda** para variables categóricas como host_is_superhost.
- Tratamiento de valores atípicos mediante **winsorización**, limitando los extremos al percentil 1% y 99% para evitar distorsión sin perder representatividad de los datos.

Análisis de correlación por tipo de habitación

Seleccione 4 tipos de habitación:

- Entire home/apt
- Private room
- Shared room
- Hotel room

Y analice los siguientes pares de variables:

1. host_acceptance_rate vs host_response_rate
2. review_scores_location vs review_scores_cleanliness
3. host_acceptance_rate vs price
4. availability_365 vs number_of_reviews
5. host_acceptance_rate vs number_of_reviews
6. reviews_per_month vs review_scores_communication

Para cada par se generaron gráficos de dispersión con regresión lineal, donde observe relaciones generalmente positivas pero moderadas entre las variables. Por ejemplo, en habitaciones tipo “Entire home/apt” se encontró una ligera relación entre limpieza y ubicación, mientras que en “Hotel room” destacó una correlación notable entre aceptación y tasa de respuesta del anfitrión.

Modelo matemático con mayor correlación por tipo de habitación

Calcule un heatmap de correlación para cada tipo de habitación y se eligió el par de variables más correlacionado (excluyendo la diagonal). Luego, se construyó un modelo de regresión lineal simple para ese par.

Tabla de las 10 correlaciones más altas (por tipo de habitación)

Para cada tipo de habitación, cree una tabla con los **10 pares de variables** más correlacionadas. A modo de ejemplo, los 5 pares más correlacionados en “Entire home/apt” fueron:

Modelos de regresión lineal múltiple

Construí un modelo de **regresión lineal múltiple** para cada una de las siguientes variables cuantitativas como variable dependiente:

- host_id
- host_acceptance_rate
- host_is_superhost
- host_total_listings_count
- accommodates
- bedrooms
- price
- review_scores_value
- reviews_per_month

El mejor modelo en cuanto a capacidad predictiva (R^2) fue el de **price**, que arrojó un coeficiente de determinación de **$R^2 = 0.46$**. Los principales predictores fueron:

Esto sugiere que el precio depende significativamente de la capacidad, número de habitaciones, ubicación, limpieza, y tipo de habitación.