Campus Puebla

Curso:

Gestión de proyectos de plataformas tecnológicas (Gpo 201)

Actividad:

Actividad 1 (Regresión Lineal Simple y Múltiple)

Elaborado Por:

Samantha Mayrin Martínez Balbuena A01733837

Fecha:

04/10/2025

Introducción

En este código se aplicó una regresión lineal simple usando la base de datos de *Airbnb Creta*, con el objetivo de analizar cómo se relacionan distintas variables como el precio, las reseñas, la disponibilidad o la tasa de aceptación del anfitrión. A lo largo del código se fueron generando gráficas y cálculos que ayudaron a identificar patrones entre los diferentes tipos de habitación (*Entire home/apt*, *Private room*, *Shared room* y *Hotel room*). Con esto se buscó entender mejor qué factores pueden influir en la valoración de los alojamientos y cómo cambian las relaciones entre variables dependiendo del tipo de espacio.

Acciones de preprocesamiento: Nulos y Outliers

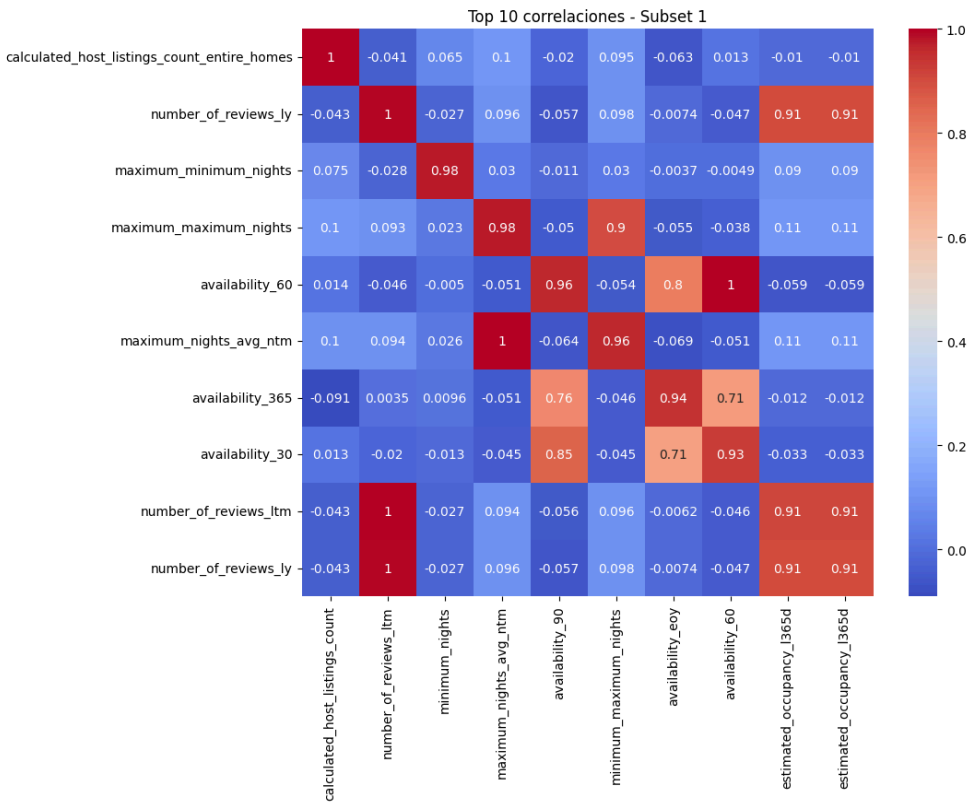
Se utilizó la base de datos listings.cvs.gz , primero se eliminó las columnas que no eran valiosas para el análisis, por ejemplo: fechas, url y id. Después se convirtieron a número la columnas que tenían caracteres especiales ($,%) y se imputaron los nulos por métodos como media, mediana, bfill, ffill, etc. Además se trataron los outliers utilizando el método de rango intercuartílico y con esto se obtuvo un base de datos limpia.

Regresión lineal simple

Se analizó la correlación que existe en cada tipo de habitación y estas fueron las 10 variables con mayor correlación.

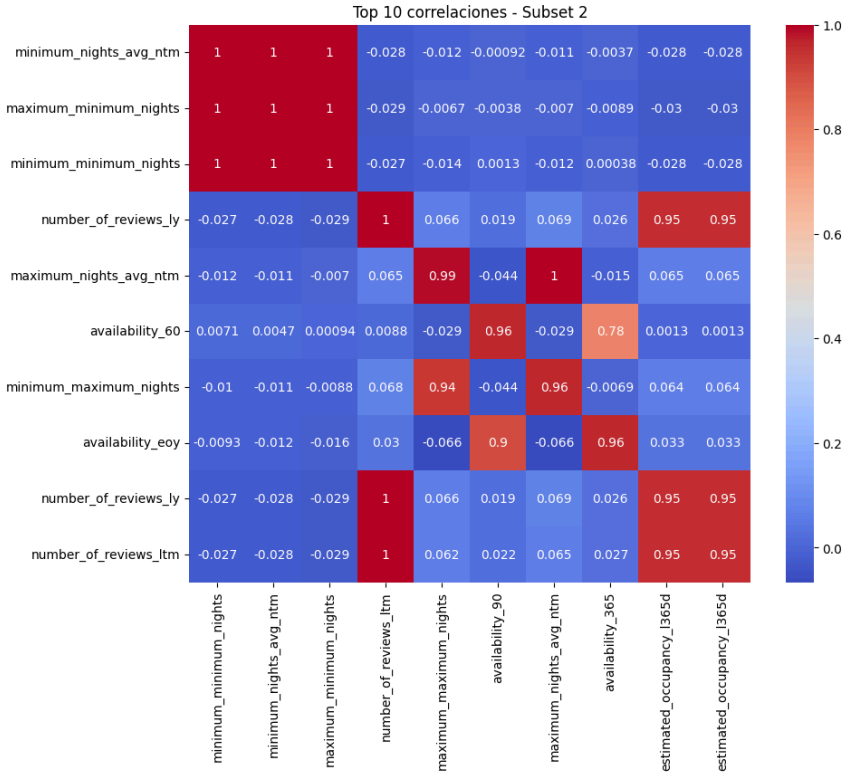
Entire home/apt

Variable_1	Variable_2	Abs_Correlación
estimated_occupancy_l365d	number_of_reviews_ltm	0.907964
estimated_occupancy_l365d	number_of_reviews_ly	0.905570
maximum_maximum_nights	minimum_maximum_nights	0.898960
bedrooms	accommodates	0.891511
host_total_listings_count	host_listings_count	0.882967
reviews_per_month	number_of_reviews_ltm	0.877823
reviews_per_month	number_of_reviews_ly	0.872806
availability_eoy	availability_90	0.857164
minimum_nights_avg_ntm	maximum_minimum_nights	0.851488
availability_90	availability_30	0.851271



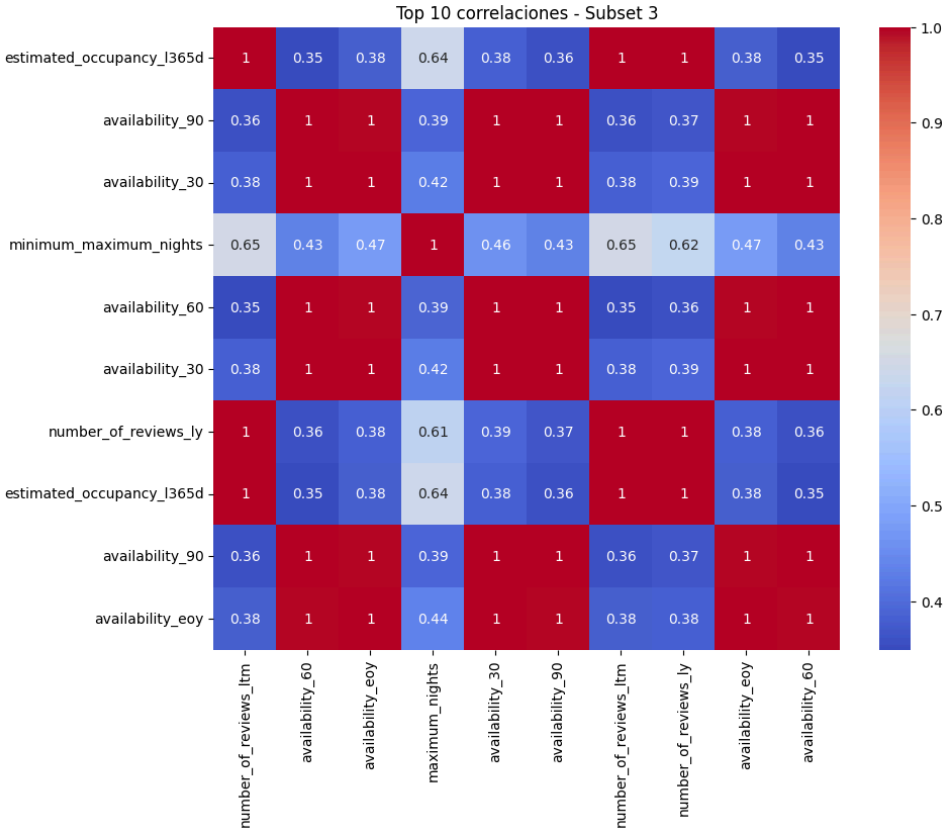
Private room

Variable_1	Variable_2	Abs_Correlación
maximum_maximum_nights	minimum_maximum_nights	0.938832
host_total_listings_count	host_listings_count	0.924655
availability_60	availability_30	0.920311
availability_eoy	availability_90	0.902147
reviews_per_month	number_of_reviews_ltm	0.887777
reviews_per_month	number_of_reviews_ly	0.883158
calculated_host_listings_count_private_rooms	calculated_host_listings_count	0.873390
review_scores_value	review_scores_rating	0.846895
reviews_per_month	estimated_occupancy_l365d	0.837890
availability_eoy	availability_60	0.837886



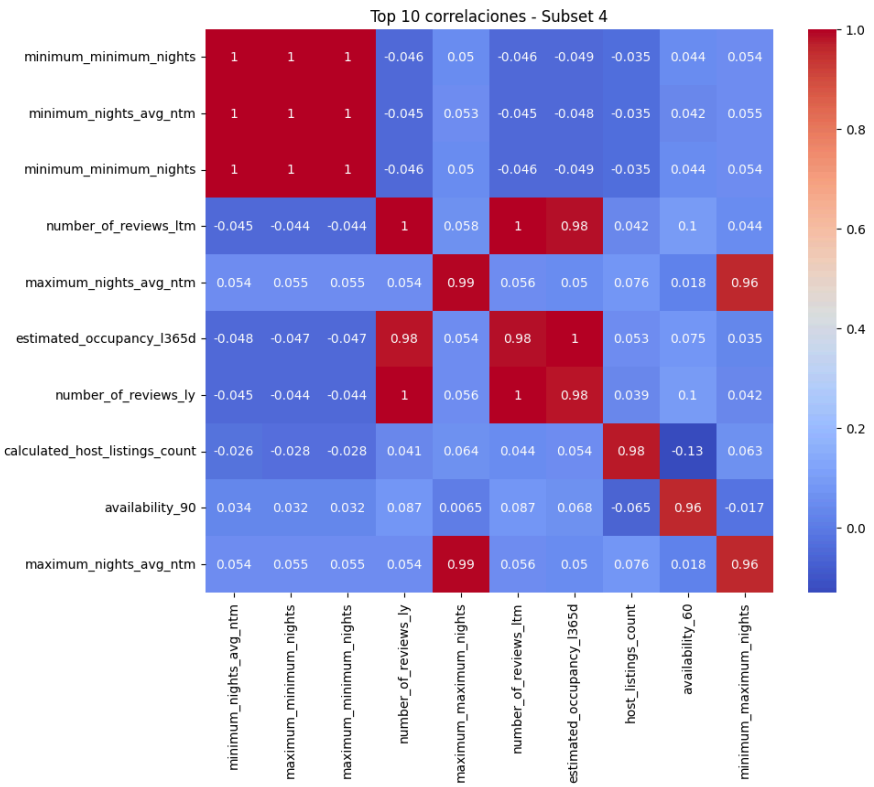
Shared room

Variable_1	Variable_2	Abs_Correlación
availability_90	availability_30	0.997764
number_of_reviews_ly	number_of_reviews_ltm	0.997534
estimated_occupancy_l365d	number_of_reviews_ly	0.997534
availability_eoy	availability_90	0.997438
availability_eoy	availability_60	0.997417
calculated_host_listings_count	host_listings_count	0.993265
availability_eoy	availability_365	0.987082
availability_365	availability_30	0.984784
availability_365	availability_60	0.981636
availability_365	availability_90	0.980156



Hotel room

Variable_1	Variable_2	Abs_Correlación
availability_eoy	availability_365	0.958711
maximum_maximum_nights	minimum_maximum_nights	0.919360
reviews_per_month	number_of_reviews	0.918408
availability_60	availability_30	0.902020
host_total_listings_count	host_listings_count	0.871527
calculated_host_listings_count	host_total_listings_count	0.847570
availability_eoy	availability_90	0.832487
availability_90	availability_30	0.804809
reviews_per_month	number_of_reviews_ly	0.804685
reviews_per_month	number_of_reviews_ltm	0.803232



Regresión múltiple

Se crea el mejor modelo de regresión lineal múltiple para cada variable cuantitativa, se eligen las variables independientes con base en las que tengan la mayor correlación con la variable independiente.

Review scores rating

```
Vars_Indep=df[['review_scores_value','review_scores_accuracy','review_scores_cleanliness']]
Var_Dep= df['review_scores_rating']
```

R = 0.746

Existe una **relación fuerte**, lo que significa que las variables de valor, precisión y limpieza influyen mucho en la calificación general del alojamiento.

Host acceptance rate

```
Vars_Indep= df[['host_response_rate', 'estimated_occupancy_l365d','number_of_reviews_ltm']]
Var_Dep= df['host_acceptance_rate']
```

R= 0.112

La relación es **muy débil**. Las variables elegidas casi no explican la tasa de aceptación del anfitrión, ya que esta depende más de su comportamiento o políticas personales.

Host is superhost

```
Vars_Indep= df[['estimated_occupancy_l365d', 'number_of_reviews_ltm','number_of_reviews_ly']]
Var_Dep= df['host_is_superhost']
```

R= 0.145

Es una **relación débil**, las variables analizadas no reflejan de manera clara quién es superhost; probablemente influyen factores como la antigüedad o las reseñas positivas.

Host total listings count

Vars_Indep= df[['host_listings_count',

'calculated_host_listings_count','calculated_host_listings_count_entire_homes']]

Var_Dep= df['host_total_listings_count']

R = 0.792

Presenta una **relación fuerte**. Las variables elegidas están muy relacionadas entre sí, ya que todas miden el número total de alojamientos gestionados por el anfitrión.

Accommodates

Vars_Indep= df[['bedrooms', 'beds','calculated_host_listings_count_entire_homes']]

Var_Dep= df['accommodates']

R = 0.787

La relación es **fuerte**, lo que indica que el número de camas y recámaras explica bien cuántos huéspedes puede recibir un alojamiento.

Bedrooms

Vars_Indep= df[['accommodates', 'beds','bathrooms']]

Var_Dep= df['bedrooms']

R = 0.770

Se observa una **relación fuerte**, ya que el número de recámaras está directamente ligado a la cantidad de camas, huéspedes y baños del anuncio.

Price

Vars_Indep= df[['host_total_listings_count', 'host_listings_count','bedrooms']]

Var_Dep= df['price']

R = 0.063

La relación es **nula**, las variables no explican el precio. Probablemente el valor depende de otros factores como la ubicación, la temporada o los servicios adicionales.

Review scores value

Vars_Indep= df[['review_scores_rating', 'review_scores_accuracy','review_scores_cleanliness']]

Var_Dep= df['review_scores_value']

R = 0.679

Hay una **relación moderada-fuerte**. El valor percibido del alojamiento se asocia con las calificaciones generales y de limpieza.

Bathroom

Vars_Indep= df[['bedrooms', 'accommodates','beds']]

Var_Dep= df['bathrooms']

R = 0.102

La relación es **muy débil**, lo que indica que el número de baños no depende directamente del tamaño o cantidad de camas.

Reviews per month

Vars_Indep= df[['number_of_reviews_ltm', 'number_of_reviews_ly','estimated_occupancy_l365d']]

Var_Dep= df['reviews_per_month']

R = 0.772

- La relación es **fuerte**. Mientras mayor es la ocupación estimada y el número total de reseñas, más reseñas mensuales tiende a recibir el alojamiento.

Durante el análisis, observé que el paso más importante antes de cualquier análisis fue la **limpieza de datos**, ya que los valores nulos y atípicos podían alterar las correlaciones y resultados. Una vez depurada la base de Airbnb Creta, las regresiones lineales simples ayudaron a identificar relaciones generales entre variables, aunque en muchos casos estas resultaron **débiles o moderadas**, lo cual tiene sentido considerando que existen factores externos como la ubicación, la temporada o las políticas del anfitrión que influyen más allá de los datos.

Después, con los **heatmaps**, se confirmó que había variables muy relacionadas entre sí, lo que permitió evitar la **multicolinealidad** y elegir solo las más representativas para los modelos múltiples. En esta segunda parte, se observaron mejoras claras en variables como **reviews_per_month**, **review_scores_rating**, **accommodates** y **bedrooms**, donde el modelo logró relaciones más sólidas. Por el contrario, en variables como **price**, **host_acceptance_rate** y **host_is_superhost**, el ajuste fue bajo, reflejando que influyen otros elementos fuera del alcance de la base de datos.

En general, la **regresión múltiple** resultó más útil que la simple, ya que permitió captar mejor las variaciones entre las variables y obtener modelos más realistas. Sin embargo, también quedó claro que el mercado de Airbnb es complejo y depende de muchos factores externos

que no siempre pueden medirse, por lo que los resultados deben interpretarse como **tendencias**, más que como predicciones exactas.