

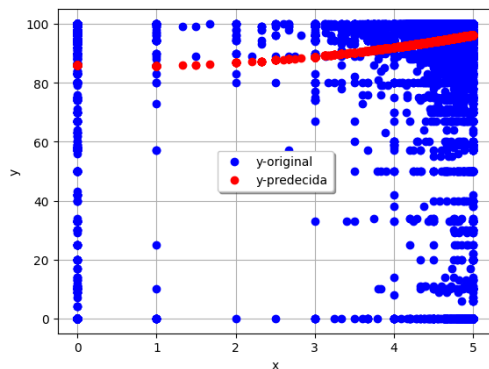
Actividad 2: Regresión No Lineal

Para esta actividad primero me aseguré que la base de datos ya no tuviera datos nulos ni valores atípicos, es importante mencionar que esto está en un código separado para evitar que los outputs tardaran mucho en cargar (en la actividad cargué la base de datos limpia y únicamente volví 2 variables numéricas para poderlas usar como independientes).

Escogí funciones cuadráticas y exponenciales ya que tienen mejores resultados en la correlación.

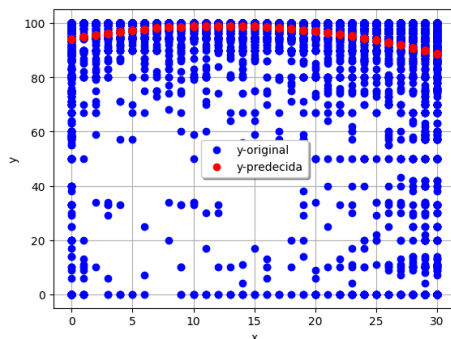
Host response rate:

Modelo 1: Host response rate vs review scores communication



Se puede observar que la mayoría de los valores se concentran en la derecha, indicando que review scores value suele ser alto, mientras que host response rate tiene un rango más amplio. Realmente no existe un patrón claro (aunque pensé que una mejor comunicación aumentaría la nota general). En el modelo lineal la correlación era de 0.189 la cual subió a 0.192 con una función cuadrática. Los valores en 0 eran datos nulos que fueron reemplazados con dicho número.

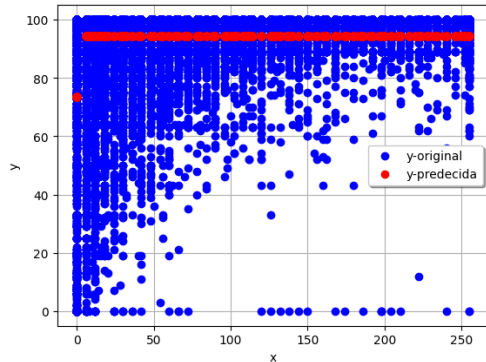
Modelo 2: Host response rate vs availability 30



En este caso tampoco se observa un patrón claro, sin embargo, las propiedades que tienen mayor disponibilidad en los 30 días son las que tienen mayor dispersión de tasa de respuesta mientras que las que menos se concentran en valores altos. La correlación original era de 0.101, la cual fue mejorada a 0.187 con una función cuadrática.

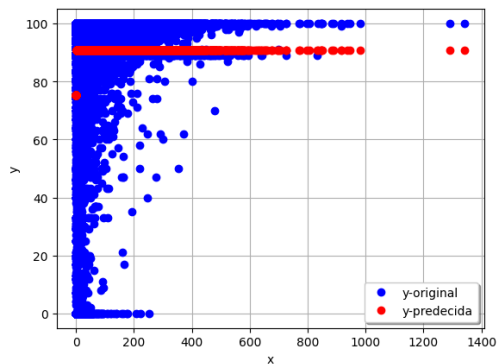
Host acceptance rate:

Modelo 1: Host acceptance rate vs estimated occupancy 1365d



Se puede observar un patrón más claro, ya que conforme aumenta la ocupación en los últimos 365 días va aumentando la tasa de aceptación, esto va de acuerdo a la lógica ya que estos anfitriones reciben a más huéspedes. En el modelo lineal la correlación era de 0.318, la cual subió a 0.396 con una función exponencial.

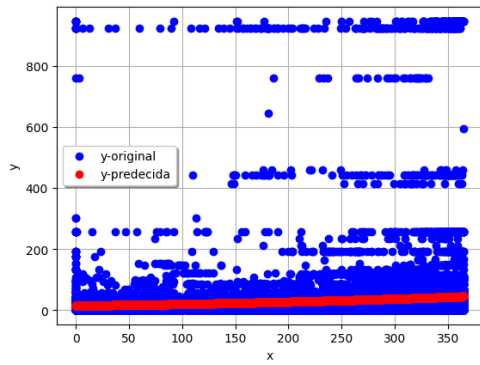
Modelo 2: Host acceptance rate vs number of reviews



En general el número de reseñas se concentra a la izquierda, aunque los anfitriones que aceptan más reservaciones tienen más que los demás, lo cual tiene sentido ya que tienen más huéspedes. La correlación lineal era de 0.198, la cual mejoró a 0.225 usando una función exponencial.

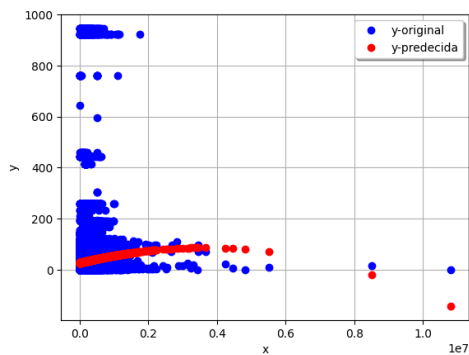
Host total listings count:

Modelo 1: Host total listings count vs availability 365



No se nota un patrón claro ya que anfitriones con un número distinto de propiedades están en todo el rango, es decir, pueden tener poca o mucha disponibilidad. Esta variable me pareció interesante ya que pensé que si tenían más propiedades habría más desocupación, sin embargo la correlación lineal era de 0.092 la cual pasó a 0.093 con una función cuadrática.

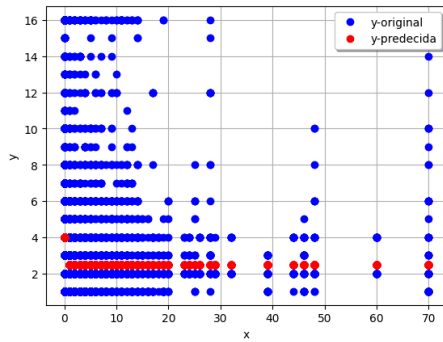
Modelo 2: Host total listings count vs estimated revenue l365d



La mayoría de propiedades se concentran a la izquierda, lo que muestra que aunque un anfitrión tenga varias esto no aumenta necesariamente el ingreso de cada una. La correlación pasó de 0.049 en el modelo lineal a 0.057 gracias a una función cuadrática.

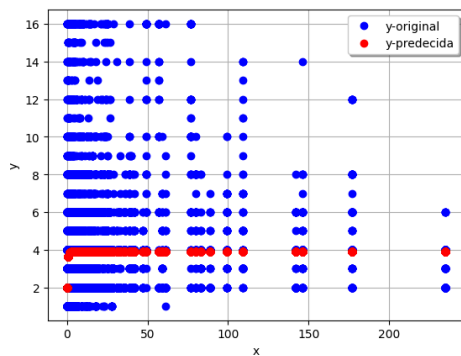
Accommodates:

Modelo 1: Accommodates vs calculated host listings count private rooms



En este caso no hay un patrón claro, es decir que aunque un anfitrión tenga más cuartos privados no necesariamente aumenta su capacidad de recibir huéspedes. La correlación lineal de 0.109 mejoró significativamente a 0.322 con una función exponencial.

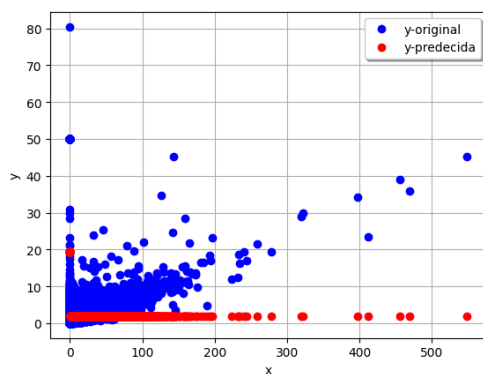
Modelo 2: Accomodates vs calculated host listings count entire homes



Aquí tampoco hay un patrón muy claro, sin embargo, se alcanza a observar como los huéspedes con más casas completas empiezan a poder recibir menos y menos huéspedes en cada una de ellas, esto significa que probablemente sean menos espectaculares cuando la oferta es mayor, enfocándose en el volumen de listings. La correlación pasó de 0.102 en el lineal a 0.340 con una función exponencial.

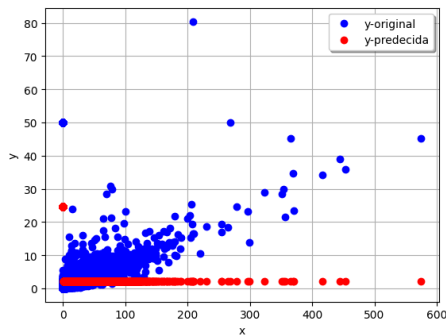
Reviews per month:

Modelo 1: Reviews per month vs number of reviews ly



Se puede observar un patrón muy claro que coincide con la lógica, ya que más reseñas el año pasado implica que se estén recibiendo más cada mes (a menos que haya un cambio muy fuerte en la tendencia). La correlación pasó de 0.157 en el lineal a 0.507 gracias a una función exponencial, teniendo una gran mejora.

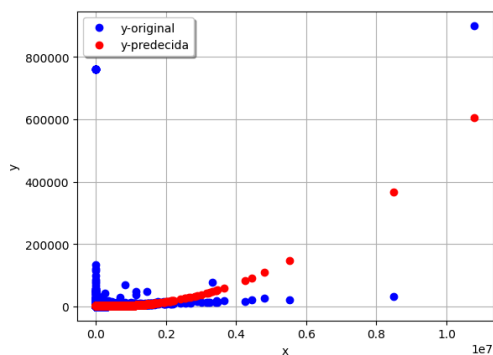
Modelo 2: Reviews per month vs number of reviews ltm



Esta variable es similar a la de number of reviews ly pero en este caso se toman en cuenta los últimos 12 meses y no los del año calendario, igualmente su comportamiento es parecido. La correlación pasó de 0.162 en el lineal a 0.613 con una función exponencial, mejorando significativamente.

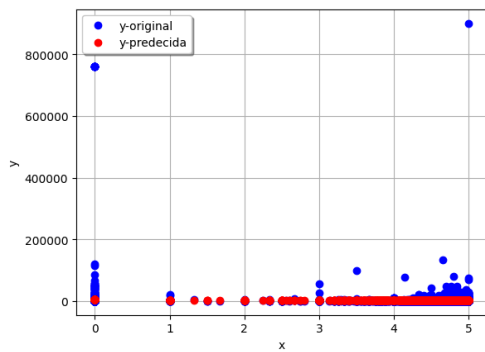
Price:

Modelo 1: Price vs estimated revenue l365d



En teoría cuando aumenta el ingreso en los últimos 365 días, uno de los factores podría ser un precio mayor, sin embargo, aunque si pasa la tendencia es muy poco marcada. La correlación pasó de 0.110 en el modelo lineal a 0.271 con una función cuadrática-

Modelo 2: Price vs review scores value



Prácticamente no existe correlación entre el precio y la calificación que los huéspedes le dan el valor, aunque si la calificación es más baja se podría pensar que fue porque el precio era caro y no cumplió las expectativas. La correlación pasó de 0.64 en el lineal a 0.66 con una función cuadrática.

	R2 (determinación)	R (correlación)
Host response rate vs review scores communication	0.036 (lineal = 0.035)	0.192 (lineal=0.189)
Host response rate vs availability 30	0.348 (lineal = 0.010)	0.187 (lineal = 0.101)
Host acceptance rate vs estimated occupancy l365d	0.156 (lineal = 0.101)	0.395 (lineal = 0.317)
Host acceptance rate vs number of reviews	0.056 (lineal = 0.039)	0.225 (lineal = 0.198)
Host total listings count vs availability 365	0.009 (lineal = 0.0086)	0.093 (lineal = 0.092)
Host total listings count vs estimated revenue l365d	0.003 (lineal =	0.057 (
Accommodates vs calculated host listings count private rooms	0.104 (lineal = 0.012)	0.322(lineal = 0.109)
Accommodates vs calculated host listings count entire home	0.116 (lineal = 0.105)	0.340 (lineal = 0.102)
Reviews per month vs number of reviews ly	0.257 (lineal = 0.025)	0.507 (lineal = 0.157) exponencial
Reviews per month vs number of reviews ltm	0.376 (lineal = 0.264)	0.613 (lineal = 0.162)
Price vs estimated revenue l365d	0.073 (lineal = 0.012)	0.271 (lineal = 0.110)
Price vs review scores value	0.043 (lineal = 0.040)	0.066 (lineal = 0.064)

En general, puedo concluir que los modelos cuadráticos lograron mejorar a los lineales en todos los casos, sin embargo, considero que la base no tiene la mejor información para realizar predicciones de unas variables con otras, especialmente al usar una sola variable independiente. Igualmente, los datos nulos que tenía afectan los resultados.

Referencias:

Get the Data. (s.f.). Inside Airbnb. <https://insideairbnb.com/get-the-data/>