



INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY

Actividad 2

José Francisco Vera Jimenez

En esta práctica se aplicaron modelos de correlación no lineal utilizando la base de datos de Airbnb.

En la actividad anterior ya se había realizado toda la parte de limpieza, depuración y eliminación de valores atípicos, por lo que en esta ocasión se trabajó directamente con los datos ya preparados.

El propósito del ejercicio fue analizar si existen relaciones no lineales entre el precio de los alojamientos (price) y otras variables del conjunto de datos, con el fin de observar comportamientos que no se ajusten a una línea recta.

A partir de la revisión de la matriz de dispersión, se identificaron algunas relaciones que presentaban formas curvas o con tendencia no lineal. En este caso, se decidió trabajar con dos de ellas: el precio en función del número de personas que puede alojar el hospedaje (accommodates) y el precio en función de las reseñas por mes (reviews_per_month).

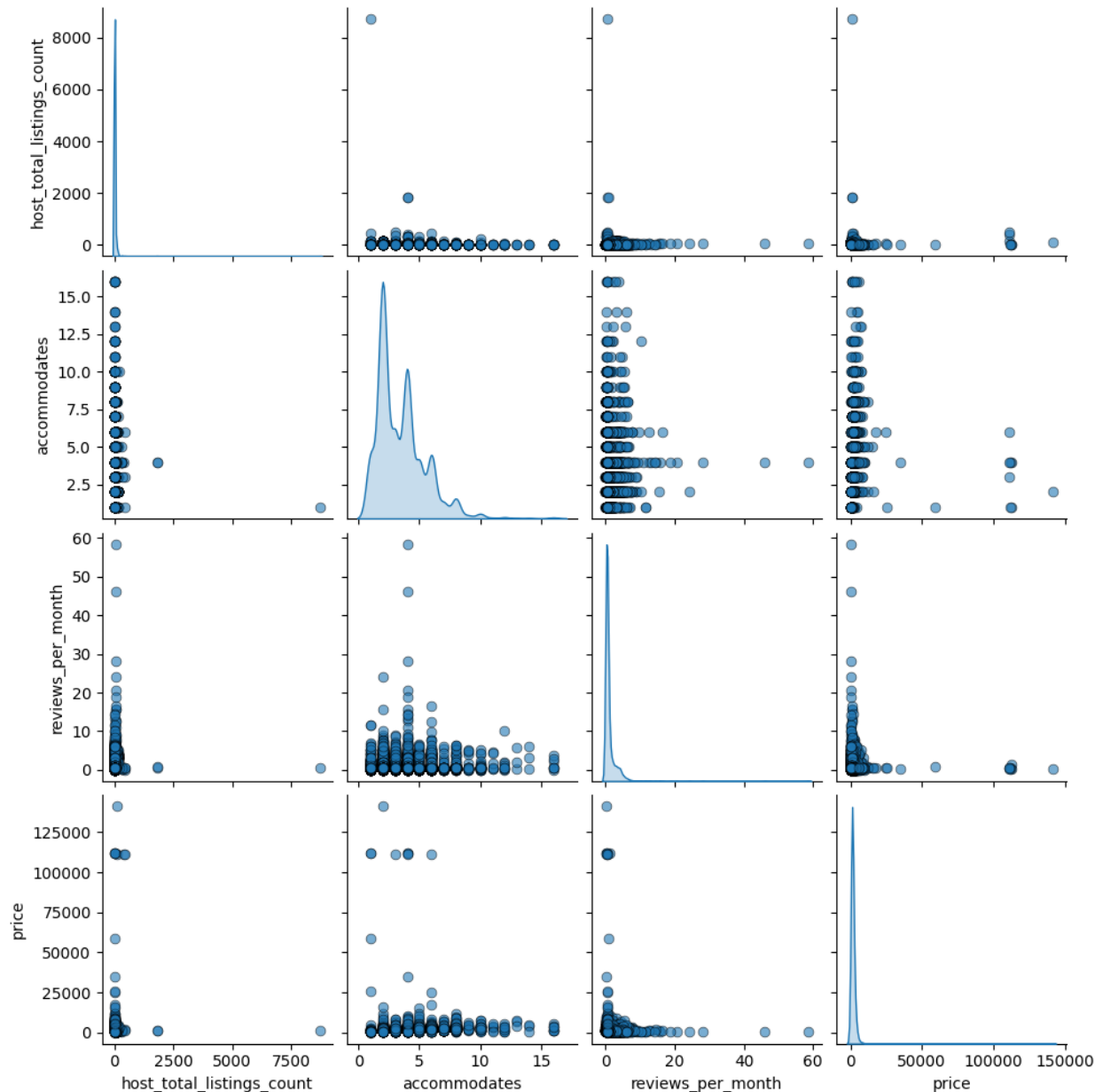
Para ambos casos se aplicaron funciones no lineales a través del método `curve_fit` de la librería SciPy, con el objetivo de ajustar los datos y obtener los parámetros de cada modelo. Posteriormente se calcularon las predicciones, así como los coeficientes de determinación (R^2) y de correlación (R), para evaluar qué tan bien se ajusta cada ecuación a los datos observados.

Al observar la matriz de dispersión entre las variables principales, se notaron algunas relaciones que no siguen un patrón lineal claro. En especial, el gráfico entre price y accommodates muestra una tendencia ascendente en forma de curva, lo que sugiere que el precio tiende a aumentar conforme crece la capacidad de alojamiento.

También se aprecia que la relación entre price y reviews_per_month es más dispersa y con una forma diferente, sin una línea definida, por lo que podría ajustarse a un modelo distinto al lineal o incluso no presentar una correlación fuerte.

A partir de esta observación se decidió aplicar dos modelos de regresión no lineal: uno cuadrático, para el caso de price y accommodates, y otro de cociente entre polinomios, para el caso de price y reviews_per_month. De esta forma se busca comparar cuál de los dos ofrece un mejor ajuste a los datos y una relación más coherente con el comportamiento real de las variables.

Matriz de dispersión entre variables (para detectar relaciones no lineales)



Para el primer modelo, se analizó la relación entre el precio (`price`) y el número de personas que puede alojar un hospedaje (`accommodates`) aplicando una función cuadrática. Con este ajuste se obtuvieron los siguientes coeficientes: **$a = -10.33$, $b = 334.35$ y $c = 784.29$** .

Estos valores indican que el precio tiende a aumentar conforme crece la capacidad del alojamiento, aunque el término cuadrático negativo sugiere que, a partir de cierto punto, el incremento deja de ser tan pronunciado. En otras palabras, el precio sube con el número de huéspedes, pero la relación no es completamente lineal, sino que se estabiliza en niveles más altos.

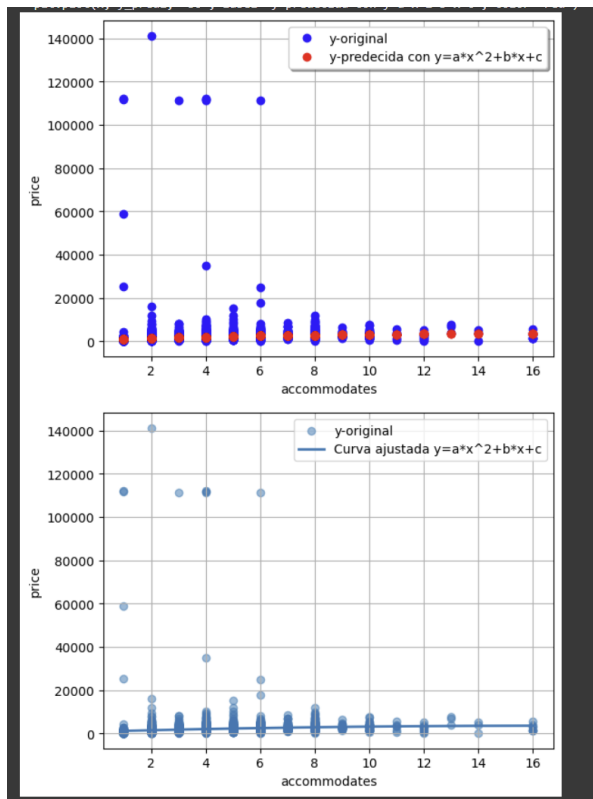
En la tabla de predicciones generada se observan valores de precio estimado que van desde alrededor de 1,100 hasta poco más de 2,100, lo que coincide con el rango general de los datos originales. La gráfica comparativa entre los valores reales y los predichos muestra que el modelo logra seguir la tendencia principal, confirmando que la relación entre ambas variables tiene un comportamiento curvo moderado y que el ajuste cuadrático describe adecuadamente dicha forma.

```
Coeficientes del Modelo 1 (a, b, c):  
[-10.33554042 334.35173269 784.29823131]  
  
Tabla de predicciones del Modelo 1:  
accommodates    y_pred1  
0              5  2197.668384  
1              2  1411.659535  
2              2  1411.659535  
3              1  1108.314424  
4              2  1411.659535  
5              1  1108.314424  
6              4  1956.336515  
7              4  1956.336515  
8              4  1956.336515  
9              2  1411.659535  
10             4  1956.336515  
11             2  1411.659535  
12             1  1108.314424  
13             6  2418.329172  
14             1  1108.314424
```

Durante la aplicación del modelo cuadrático se observó que los valores de precio (price) presentaban una gran dispersión, con algunos casos muy altos que afectan la forma general de la curva. Estos valores extremos, aunque no fueron eliminados como *outliers* en la limpieza anterior, sí provocan que el modelo se concentre en los precios más bajos y no logre representar correctamente los casos con montos más elevados.

Para reducir este efecto sin eliminar datos, se decidió realizar una segunda visualización aplicando una curva de suavizado, lo que permitió observar de forma más clara la tendencia general del modelo. En esta gráfica se aprecia que la relación entre el número de personas que puede alojar un hospedaje (accommodates) y el precio mantiene una forma creciente, aunque el aumento deja de ser tan pronunciado a partir de cierto punto.

En conjunto, los resultados muestran que el precio tiende a incrementarse conforme crece la capacidad del alojamiento, pero con un límite natural donde el incremento deja de ser proporcional. El uso de la curva suavizada ayudó a visualizar mejor esta tendencia general sin que los valores extremos distorsionaran el comportamiento principal de los datos.



El modelo cuadrático aplicado a la relación entre price y accommodates arrojó un coeficiente de determinación (R^2) de 0.0100 y un coeficiente de correlación (R) de 0.1004. Estos resultados confirman que la relación entre ambas variables es débil y que el modelo no logra explicar gran parte de la variabilidad del precio.

Aunque la forma de la curva se ajusta a la tendencia general donde el precio tiende a aumentar conforme crece la capacidad de alojamiento, la dispersión de los datos y la presencia de precios muy altos hacen que el modelo no tenga un buen desempeño estadístico.

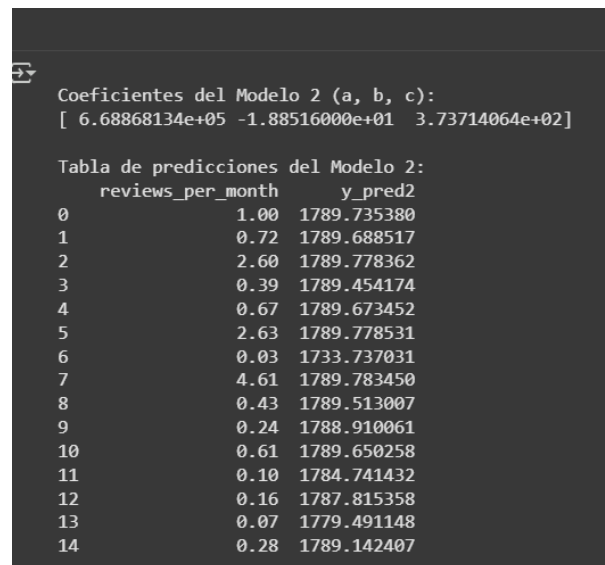
Aun así, la curva suavizada permitió visualizar con mayor claridad el comportamiento general: una tendencia ascendente moderada que refleja que, en promedio, los alojamientos con mayor capacidad suelen tener precios más altos, aunque con una relación no proporcional ni constante.

Para el segundo modelo:

Se utilizó la función Cociente entre Polinomios, con la relación entre price (variable dependiente) y reviews_per_month (variable independiente). Este modelo se eligió para evaluar si el precio podía estar influenciado de manera no lineal por la frecuencia de reseñas mensuales de los alojamientos.

Los parámetros obtenidos fueron $a = 6.6886813e+05$, $b = -18.5160000$, y $c = 373.714064$, lo que indica que el numerador del polinomio tiene un crecimiento cuadrático positivo, mientras que el denominador atenúa la variación del precio conforme aumentan las reseñas. En otras palabras, el modelo sugiere una relación inversa suave: los alojamientos con pocas reseñas tienden a mostrar precios más elevados, mientras que conforme aumenta el número de reseñas mensuales el precio se estabiliza alrededor de un rango constante.

La tabla de predicciones muestra valores relativamente uniformes, concentrados alrededor de los 1789–1790, lo que confirma que la influencia de reviews_per_month sobre el precio es limitada dentro del rango de datos analizado. Esto sugiere que, aunque el modelo es válido matemáticamente, la variable de reseñas mensuales no presenta un impacto fuerte en la determinación del precio.



```
↵
Coeficientes del Modelo 2 (a, b, c):
[ 6.68868134e+05 -1.88516000e+01  3.73714064e+02]

Tabla de predicciones del Modelo 2:
reviews_per_month  y_pred2
0                1.00  1789.735380
1                0.72  1789.688517
2                2.60  1789.778362
3                0.39  1789.454174
4                0.67  1789.673452
5                2.63  1789.778531
6                0.03  1733.737031
7                4.61  1789.783450
8                0.43  1789.513007
9                0.24  1788.910061
10               0.61  1789.650258
11               0.10  1784.741432
12               0.16  1787.815358
13               0.07  1779.491148
14               0.28  1789.142407
```

Al comparar ambos modelos no lineales aplicados a la variable price, se observa que el comportamiento general de los datos presenta una alta concentración de valores en precios bajos, mientras que existen pocos registros con precios muy elevados que generan gran dispersión.

En el modelo cuadrático ($y = a \cdot x^2 + b \cdot x + c$), la curva ajustada muestra una tendencia creciente inicial seguida de una ligera caída, lo que refleja que el precio tiende a aumentar con el número de huéspedes (accommodates), pero no de manera proporcional. Sin embargo, los valores extremos del precio reducen considerablemente la calidad del ajuste, provocando una forma de curva que se mantiene muy cercana al eje y apenas logra representar la variabilidad total de los datos.

Por otro lado, el modelo de cociente entre polinomios ($y = (a \cdot x^2 + b) / (c \cdot x^2)$) aplicado con la variable reviews_per_month muestra una relación inversa, donde los alojamientos con pocas reseñas presentan precios más altos y, conforme aumenta el número de reseñas mensuales, el precio se estabiliza hacia valores bajos y constantes. Aunque el modelo se ajusta correctamente desde el punto de vista matemático, la dispersión y la concentración de datos en valores pequeños hacen que la curva predicha se mantenga casi horizontal.

En conjunto, ambos modelos permiten identificar tendencias generales del comportamiento del precio, pero dejan claro que esta variable depende de más factores además de las reseñas o la capacidad del alojamiento. Por tanto, aunque los ajustes no alcanzan un alto nivel de precisión, sí permiten visualizar de manera aproximada cómo varían los precios dentro del conjunto de datos analizado.

En conclusión, los modelos no lineales ayudaron a entender mejor cómo se comporta el precio dentro del conjunto de datos, aunque los resultados muestran que las relaciones con variables como accommodates o reviews_per_month no son tan fuertes como se esperaba.

El modelo cuadrático mostró una ligera tendencia al alza en los precios conforme aumenta la capacidad del alojamiento, mientras que el modelo de cociente entre polinomios reflejó una relación inversa con las reseñas mensuales. Sin embargo, la alta variación de los precios y la presencia de valores extremos limitaron el ajuste de ambos modelos.

Aun con ello, el ejercicio permitió visualizar que el comportamiento del precio no sigue una línea recta y que, en este tipo de datos, las relaciones suelen ser más complejas y requieren modelos no lineales para poder apreciarse con mayor claridad.

Conclusión:

Los modelos no lineales ayudaron a entender mejor cómo se comporta el precio dentro del conjunto de datos, aunque los resultados muestran que las relaciones con variables como `accommodates` o `reviews_per_month` no son tan fuertes como se esperaba.

El modelo cuadrático mostró una ligera tendencia al alza en los precios conforme aumenta la capacidad del alojamiento, mientras que el modelo de cociente entre polinomios reflejó una relación inversa con las reseñas mensuales. Sin embargo, la alta variación de los precios y la presencia de valores extremos limitaron el ajuste de ambos modelos.

Aun con ello, el ejercicio permitió visualizar que el comportamiento del precio no sigue una línea recta y que, en este tipo de datos, las relaciones suelen ser más complejas y requieren modelos no lineales para poder apreciarse con mayor claridad.