Ciencia de Datos

Trabajo Práctico N° 2

Profesora: Maria Noelia Romero

Tutor: Ignacio Spiousas

Gabriel Alejandro Díaz y Gabriela Belen Sanchez

2 de octubre de 2024

Limpieza de Datos

En esta sección se detalla el proceso de limpieza de la base de datos que contiene a los oferentes de Airbnb localizados en la ciudad de Nueva York, Estados Unidos. De esta forma, se busca garantizar la integridad y precisión de los datos, de forma tal que el posterior análisis y predicciones derivadas de las estas no se vean afectadas.

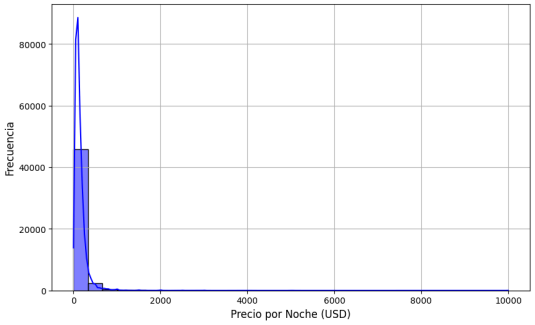
Inicialmente, la base de datos contaba con un total de 48.905 registros de anuncios, organizados en múltiples columnas con atributos de la propiedad publicada. Específicamente, estas variables consisten en: "id" (identificador único para cada anuncio), "name" (nombre o titulo del anuncio), "host_id" (Hid, identificador del anfitrión), "host_name" (nombre del anfitrión), "neighbourhood_group" (Neg, barrio o ubicación del alojamiento), neighbourhood (Ne, barrio más detallado), Latitud, Longitud, "room_type" (Rt, tipo de alojamiento disponible: *Private room, Entire home/apt, Shared room*), price (precio por noche del anuncio), "minimum_nights" (Mn, número de noches mínimo requerido para una reserva), "number_of_reviews" (NoF, número total de reseñas del anuncio), "last_review" (fecha de la última reserva), "reviews_per_month" (RpM, reseñas por mes), "calculated_host_listings_count" (CHLC, número de anuncios del anfitrión en la plataforma) y "availability_365" (Av, número de días al año que el anuncio está disponible).

El proceso de limpieza se realizó mediante Python, específicamente con los módulos *Pandas* y *Numpy*. Este comenzó con la verificación de registros duplicados en la base. Específicamente, se encontraron 10 registros duplicados con respecto a todas sus características, por lo que se procedió a su eliminación (n = 48.895). Posteriormente, se descartaron aquellas variables que no presentaban información útil para los objetivos predictivos del presente trabajo. De esta forma, se optó por trabajar únicamente con los atributos: Neg, Ne, Lat, Long, RT, Precio, Mn, NoR, RpM, CHLC y Av.

A continuación, se procedió a buscar posibles *missing values* dentro de la base de datos. Se encontraron un total de 10.067 datos faltantes, de los cuales 10.052 se encontraban en la variable RpM y 15 en Precio. Para su tratamiento, se optó por analizarlos de forma separada según su variable. En un primer lugar, al analizar detenidamente los registros con datos faltantes en RpM, se observó que todos presentaban NoR = 0, es decir, contaban con 0 reviews totales. De esta forma, se concluyó que esta *missing data* se podía deber a que, en el proceso de recolección de la información, para todos aquellos registros con cero reviews no se realizó el cálculo de RpM. En cambio, se dejó el espacio en blanco. En otras palabras, se trataba de data *Missing Not at Random* (Gelman y Hill, 2006), existe una relación entre los datos faltantes y sus valores. En base a esto, para su solución, simplemente se rellenó con un "0" aquellas casillas con un 0 en NoR. Para corroborar su correcto funcionamiento, luego de este proceso se volvió a identificar los posibles *missing values*, y solo se encontraron aquellos 15 pertenecientes a precio.

Figura 1

Densidad de los precios de oferentes de Airbn en Nueva York



Nota. El gráfico muestra la frecuencia de los precios por noche de oferentes de Airbnb en Nueva York.

Al analizar los datos de faltantes de esta última categoría, se observó que no presentan una distribución lineal y que existe un amplio rango de valores que este puede tomar, con una mayor distribución de valores cercanos a los 200 USD. En base a esta información, se optó por realizar una imputación de los datos faltantes utilizando la mediana de los precios. Este método se trata de uno de los más fáciles de realizar y permite, al utilizar la mediana, acoplarse a una mayor cantidad de distribuciones en comparación con la media. Aún así, este método también conlleva sus desventajas: En primer lugar, si el número de *missing data* es muy amplio, es posible que, por la imputación, la distribución de los datos se vea afectada, se reduce la desviación estándar conforme aumenta esta cantidad (Anil, et al., 2019). A su vez, puede distorsionar las relaciones entre las variables al modificar los estimadores de correlación (Gelman y Hill, 2006), y, en general, puede presentar estimaciones altamente sesgadas (Anil, et al., 2019), aunque este último puede reducirse debido a que la mediana se trata de un estimador menos afectado por los valores extremos (Pham-Gia y Hung, 2001). Aún así, se optó por esta metodología debido a que estudios han señalado que las limitaciones de este tipo de imputación se reducen si la *missing data* es menor al 10% de los datos totales (en este caso es un 0.030%) (Anil, et al., 2019) y se estratifican los datos en distintos subgrupos (Gelman y Hill, 2006). Por esta razón, se decidió realizar la imputación con las medianas de los subgrupos conformados por aquellos anuncios que comparten igual barrio (Ne) y tipo de alojamiento (RT) que el *missing data.*

Posteriormente se buscó la presencia de outliers u otro tipo de valor que no tuviera sentido dentro del set de datos. En primer lugar, se observó que existen 11 registros que tienen como precio registrado un 0, es decir *price* = 0. En estos casos, debido a que el enfoque del estudio es la predicción de los precios de alojamiento, se concluyó que estos no aportan información valiosa ya que podría representar algún tipo de promoción, pruebas o simplemente propiedades antiguas que no se encuentran actualmente disponibles. En consecuencia, se optó por su eliminación (n_{nueva} = 48.884). Posteriormente también se analizó a los outliers en base al MAD, teniendo en cuenta la mediana de los subgrupos mencionados anteriormente. De esta forma, se descartaron aquellos valores que sobrepasaron el umbral: mediana ± MAD x 1.5 (14.860) y se obtuvo un nuevo *n* total de registros de 34.024. Aún así, como se verá más adelante, esto también puede conllevar a que haya registros con valores altos en comparación con los demás, debido a que son pocos o únicos en su categoría tras la limpieza de datos.

Por otro lado, se observó que la variable Mn (*minimum_nights*) presenta una gran cantidad de valores altos. Específicamente, 491 registros poseían valores superiores al mes (30 días). Debido a que el enfoque de este estudio se basa en el análisis de periodos de alquiler a corto plazo, característicos de Airbnb, se optó por el descarte de estos registros (n = 33.533). De forma similar, se descartaron todos aquellos registros que poseían la variable Av (*availability_365*) igual o inferior a 0 (12.512). Esto fue así decidió ya que estos anuncios no se encuentran disponibles, por lo que la información asociada a

ellos, como los precios o número de reviews, se puede encontrar desactualizada o simplemente puede ser incorrecta (n = 21.021). El resto de las variables no fue considerado como posibles *outliers*, debido a que tiene sentido que existan algunos valores altos en las mismas.

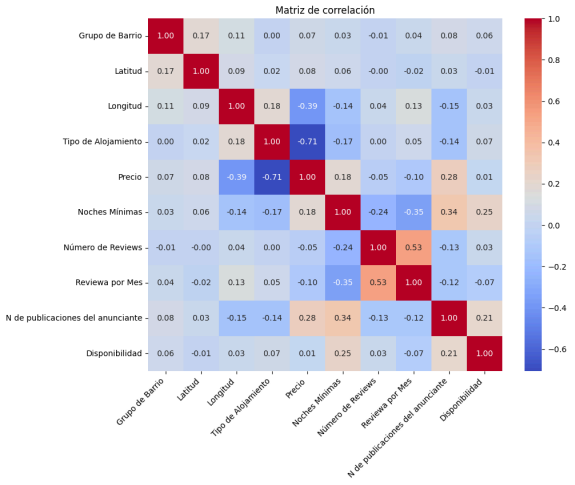
Una vez que la base de datos fue limpiada, se creó una nueva columna (*offer_group*, OG) que contiene la cantidad de oferentes por cada NeG, para futuros análisis.

Gráficos y visualizaciones

Una vez hecha la limpieza de datos, realizamos una serie de gráficos y análisis que nos permitan explorar qué relaciones se pueden establecer entre las variables. Para el ejercicio 2, construimos una matriz de correlación entre las siguientes variables: NeG, Lat, Long, RT, Precio, Mn, NoR, RpM, CHLC y Av. En este sentido, la matriz lo que calcula es el coeficiente de correlación de Pearson entre cada par de variables. Visualizamos la matriz con un *heatmap* (mapa de calor) generado por el código *sns.heatmap*, en donde los valores de correlación se anotan en cada celda del gráfico para facilitar la interpretación.

Figura 2

Matriz de correlaciones de las variables relevantes



Nota. El gráfico muestra las correlaciones entre las variables neighbourhood_group_cod, latitude, longitude, room_type_cod, price, minimum_nights, number_of_reviews, reviews_per_month, calculated_host_listings_count y availability_365 a través del coeficiente de Pearson para cada par de variables.

En base a los resultados obtenidos observamos una serie de correlaciones que resultan de interés a la hora de interpretar los datos. Por un lado, observamos que la correlación entre Precio y Tipo de Alojamiento es de -0.71 (fuerte y negativa). Esto sugiere que los precios tienen una tendencia a disminuir conforme se pasa de alojamientos de casas/departamentos (*entire home/apt*) a habitaciones privadas (*private room*) o compartidas (*shared room*) con alta significatividad. Por otro lado, vimos que la correlación entre Precio y Número de Reviews es cercana a 0 (-0.05). Esto indica que no existe una relación significativa entre el número de reseñas y el precio del alojamiento. Puede deberse a que las reviews reflejan más la popularidad que factores como el precio, y, a su vez, que los propietarios no cambian el precio según la cantidad de las reseñas. Aún queda la duda sobre cómo afectarían las

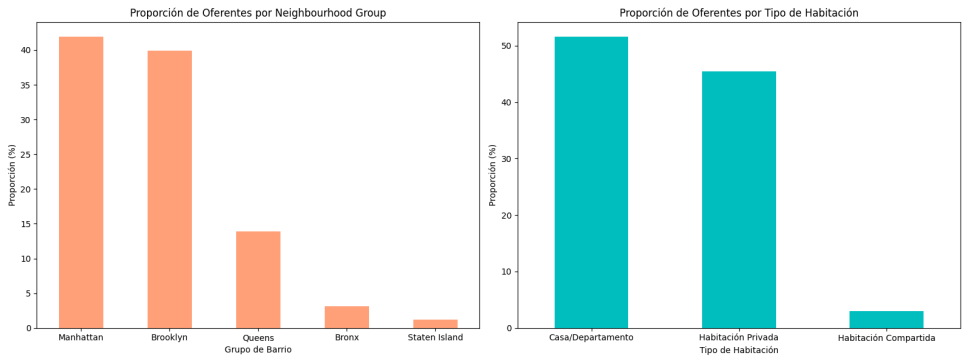
calidades (positivas o negativas) sobre este. Por otro lado, otra correlación relativamente alta es la de Precio y Longitud (-0.39), lo cual indicaría que el precio de las propiedades aumenta conforme esta se encuentra más al oeste (menor latitud) de la ciudad de Nueva York. A su vez, se vio que la correlación entre Precio y Noches Mínimas es de 0.18 (débil y positiva). Esto puede sugerir que los alojamientos con un mayor requisito de noches mínimas suelen tener precios más altos, aunque la correlación es baja. También, se observó que la correlación entre Precio y Número de Publicaciones del Anunciante es de 0.28 (positiva y débil), que indicaría que a medida que aquellos anfitriones con más propiedades suelen publicarlas a un mayor precio.

Como cabría esperarse, notamos que el Número de Reviews y Reviews por Mes tienen una correlación positiva y moderada de 0.53, ya que el segundo depende del primero en cierta medida. Por último, nos resultó relevante destacar que la correlación de 0.25 entre Disponibilidad y Noches Mínimas nos podría indicar que las propiedades con un mayor número de noches mínimas suelen estar más disponibles a lo largo del año.

En el ejercicio 3 se calculó la proporción de oferentes para cada grupo de barrio (*neighbourhood_group*), así como para cada tipo de alojamiento (*room_type*). Realizamos 2 (dos) gráficos de barras para representar visualmente estas proporciones. Por un lado, observamos que Manhattan representa el mayor porcentaje de oferentes con un 41.94%, seguido de Brooklyn con un 39.89%. Estos dos barrios concentran más del 80% de las propiedades que están Airbnb, lo que sugiere que son las zonas más populares y demandadas. En este sentido, barrios como Queens, Bronx, y Staten Island tienen una representación mucho menor, lo que podría indicar que son áreas menos atractivas para los turistas o que el mercado de Airbnb en estas zonas está menos desarrollado. Por otro lado, en relación a la proporción de oferentes para cada tipo de alojamiento, vimos que el 51.63% de las propiedades son *entire home/apt* (casas o departamentos), lo que sugiere que más de la mitad de las propiedades en Airbnb son de tipo individual y privado (no tienen que compartir con nadie y no es una habitación dentro de una casa de familia). A su vez, se observó que un 45.42% de las propiedades son *private room* (habitaciones privadas), lo que podría indicar que muchas personas alquilan habitaciones en su hogar. Finalmente, notamos que las *shared room* (habitaciones compartidas) representan solo el 2.95%, por lo que no son muy populares.

Figura 3

Proporción de oferente por neighborhood group y por tipo de habitación

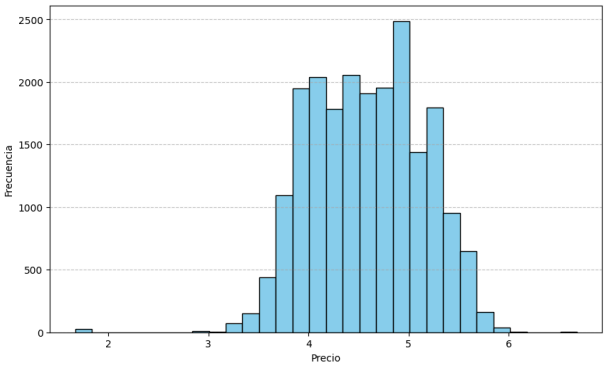


Nota. El gráfico muestra la proporción (en porcentaje) de oferentes por neighborhood group (*i.e*, en Manhattan, Brooklyn, Queens, Bronx y Staten Island) y por tipo de habitación (casa/departamento, habitación privada y habitacion compartida)

Para el ejercicio 4 creamos un histograma de los precios de los alojamientos con la variable transformada *price_log*. La transformación logarítmica la aplicamos previamente para tratar de normalizar la distribución, ya que observamos que los precios tendían a ser asimétricos, con una distribución sesgada a la derecha (muchos valores bajos y pocos valores muy altos). Con la transformación logarítmica logramos una forma más simétrica, con una mayor normalización de la variable. A su vez, vimos que la mayoría de los precios están concentrados en el rango de 3.5 a 5.8 (en términos de *price_log*), lo que en la escala original de precios corresponde a un rango de aproximadamente 30 a 300 USD por noche. El precio mínimo es de 5 USD, el máximo de 800 USD y el promedio de aproximadamente 114 USD. Los alojamientos con precios más altos elevan considerablemente el precio promedio. Por otro lado, los precios medios muestran que Manhattan es el barrio más caro con un promedio de 146.92 USD por noche, mientras que el Bronx tiene los precios más bajos con un promedio de 69.58 USD por noche. Asimismo, notamos que los alojamientos de estilo *entire home/apt* tienen un precio promedio significativamente más alto (158.45 USD) que las habitaciones privadas (68.91 USD) y compartidas (49.30 USD).

Figura 4

Histograma de los precios de alojamiento

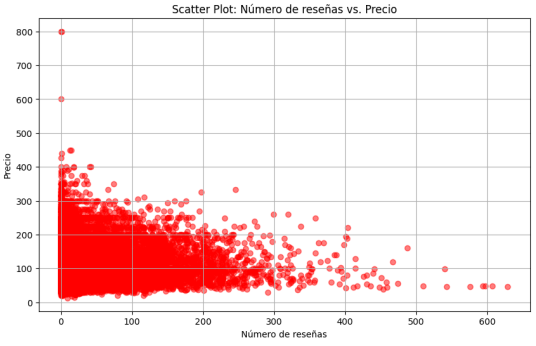


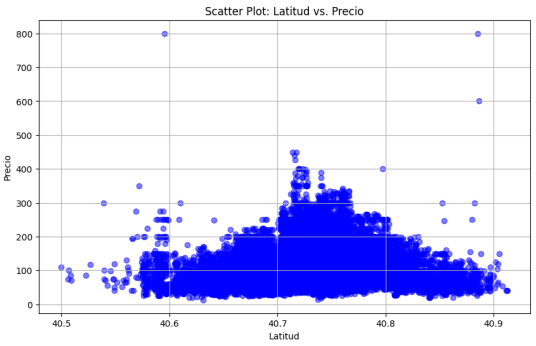
Nota. El gráfico muestra la frecuencia de los precios de alojamiento basado en la clasificación de *precio_log*.

En el ejercicio 5 creamos dos gráficos de dispersión para visualizar la relación entre *latitud* y *price*, y *number_of_reviews* y *price*. Para la relación entre latitude y price observamos que la mayor concentración de precios se encuentra en una franja delimitada entre las latitudes 40.7 y 40.8. Esto sugiere que en esta zona específica de la ciudad (Manhattan y/o Brooklyn), hay más alojamientos disponibles, concordante con la proporción de oferentes por neighborhood group. También, se observa que la variabilidad de precios en esta franja es alta, con precios que van desde 10 USD hasta 450 USD por noche. No se logró apreciar una relación lineal directa entre latitud y precio. Esto podría sugerir que la latitud por sí sola no es un buen predictor del precio, y que otras variables pueden estar influyendo en el valor de los alojamientos. Respecto a la relación de número de reviews y price observamos una tendencia decreciente: a medida que aumenta el número de reseñas, los precios tienden a disminuir. Los precios más altos (hasta 800 USD por noche) tienden a concentrarse en alojamientos con pocas reseñas (menos de 100). A su vez, vimos que la mayoría de alojamientos con precios bajos (por debajo de 100 USD) tienen un número elevado de reseñas, lo que podría indicar que son alojamientos más accesibles y/o populares.

Figura 5 y 6

Scatter Plot: Número de reseñas vs. Precio y Scatter Plot: Latitud vs. Precio





Nota.

Ambos gráficos muestran las relaciones entre las variables Precio y Número de reseñas y Precio y latitud

Para el ejercicio 6 realizamos un análisis de componentes principales (PCA) para reducir las variables a dos dimensiones (CP1 y CP2). Las variables numéricas fueron estandarizadas y luego se proyectaron en un espacio de dos componentes principales, donde se mantuvo la mayor varianza posible de los datos. Estos componentes principales explicaron un 43% de la varianza total de los datos (CP1 explicó el 26% de la varianza, mientras que el CP2 el 17%). Esto indica que, aunque se perdió información al reducir las dimensiones, aún se conserva una proporción razonable de la varianza total.

Tabla de Loadings

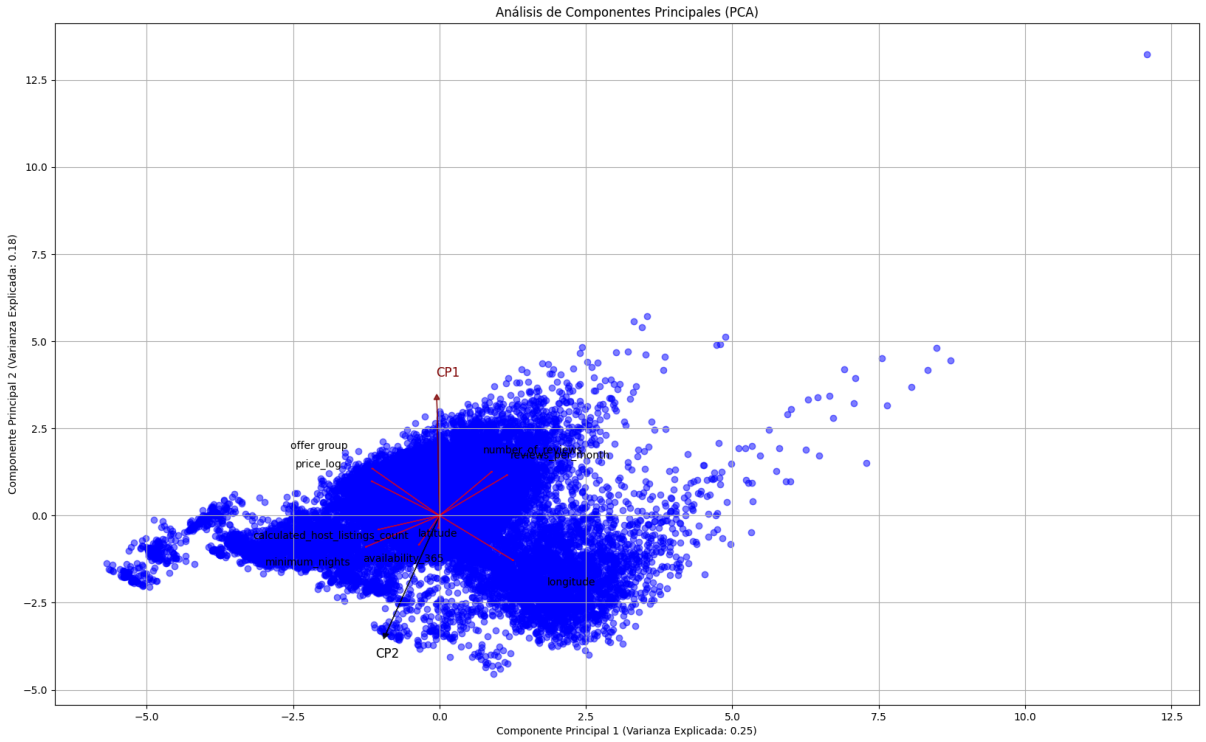
Loadings de cada variable en los componentes principales:

	PC1	PC2
latitude	-0.005659	-0.114360
longitude	0.416141	-0.423904
price_log	-0.381704	0.325499
minimum_nights	-0.415190	-0.296426
number_of_reviews	0.295013	0.416237
reviews_per_month	0.382111	0.384838
calculated_host_listings_count	-0.343375	-0.131221
availability_365	-0.115532	-0.274625
offer group	-0.380025	0.446562

En base a sus *loadings*, es decir, el aporte de cada variable al componente principal, observamos que las variables que más contribuyen al CP1 son *price_log*, *minimum_nights* y *reviews_per_month* lo que podría indicar que la primera dimensión está principalmente relacionada con factores que influyen en el precio y el comportamiento de los propietarios. Las variables longitud, latitud y *offer group* tienen una contribución importante en CP2, lo que sugiere que la segunda dimensión está más relacionada con la ubicación geográfica.

Figura 7

Análisis de componentes principales



Nota. El gráfico muestra la direcciones de primer y segundo componente principal.

Predicción

Finalmente, en esta sección del trabajo se presenta el desarrollo de un modelo predictivo para estimar el precio de los alquileres temporales anunciados en la plataforma de Airbnb. Para ello, se realizó una regresión lineal que toma como variables regresoras a todas las variables trabajadas hasta el momento (Latitud, Longitud, Noches Mínimas, Número de Reviews, Reviews por Mes, N° de publicaciones del anunciante y Disponibilidad). Cabe recalcar que, como en el caso anterior, este análisis se realizó sobre el logaritmo de la variable y (precio) debido a la distribución presentada por esta variable. De esta forma, el modelo predictivo es:

$$\text{Log}(precio) = x_0 + \beta_1 Lat + \beta_2 Long + \beta_3 Mn + \beta_4 NoF + \beta_5 RpM + \beta_6 CHLC + \beta_7 Av$$

Para la construcción del modelo se dividió a la base de datos en dos partes: un conjunto de entrenamiento (*x_train* e *y_train*), compuesto por el 70% de los datos restantes (n=14714); y un conjunto de prueba (*x_test* e *y_test*) constituido por el 30% restante (n=6307).

Resultados

OLS Regression Results

Dep. Variable:	price_log	R-squared:	0.224
Model:	OLS	Adj. R-squared:	0.223
Method:	Least Squares	F-statistic:	529.7
Date:	Wed, 02 Oct 2024	Prob (F-statistic):	0.00
Time:	13:00:34	Log-Likelihood:	-10060.
No. Observations:	14714	AIC:	2.014e+04
Df Residuals:	14705	BIC:	2.021e+04
Df Model:	8		
Covariance Type:	nonrobust		

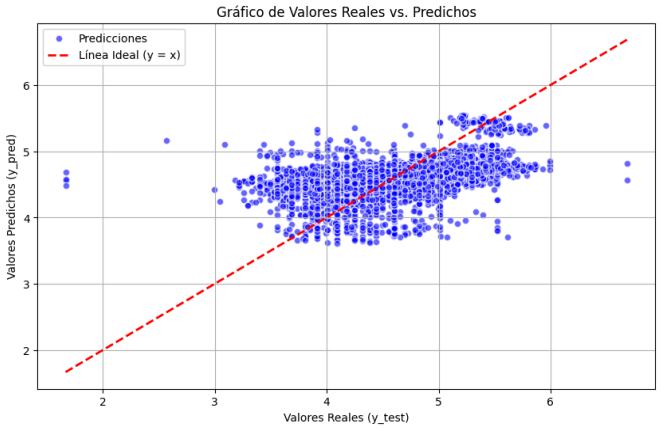
	coef	std err	t	P>|t|	[0.025	0.975]
latitude	1.0895	0.070	15.663	0.000	0.953	1.226
longitude	-2.8751	0.091	-31.587	0.000	-3.054	-2.697
minimum_nights	0.0024	0.000	5.002	0.000	0.001	0.003
number_of_reviews	0.0003	8.53e-05	3.835	0.000	0.000	0.000
reviews_per_month	-0.0038	0.003	-1.441	0.150	-0.009	0.001
calculated_host_listings_count	0.0022	0.000	20.454	0.000	0.002	0.002
availability_365	-0.0002	3.34e-05	-4.531	0.000	-0.000	-8.59e-05
offer group	3.977e-05	1.92e-06	20.694	0.000	3.6e-05	4.35e-05
intercept	-252.7012	7.478	-33.793	0.000	-267.359	-238.044

Omnibus:	189.601	Durbin-Watson:	2.014

Los resultados del modelo de regresión lineal muestran que la variable dependiente, el logaritmo del precio (***price_log***), tiene un r cuadrado (R^2) de 0.224, lo que indica que aproximadamente el 19.1% de la variabilidad del precio en la muestra de entrenamiento se explica por las variables independientes seleccionadas. Similarmente, el r cuadrado ajustado también es de 0.223, lo que confirma que el modelo no pierde mucha capacidad explicativa al ajustar por el número de predictores utilizados en este caso.

Figura 8

Gráfico de valores reales vs Predichos



Nota. El gráfico muestra la relación entre valores predichos (= valores reales) y la línea ideal (roja punteada).

Finalmente, el Error Cuadrático Medio (MSE) es de 0.23, lo cual indica la magnitud de los errores en el modelo estadístico al darnos la media de las diferencias entre los valores reales y los predichos de los β_i. En este contexto, donde se está estimando el logaritmo del precio, este error equivale a una desviación del precio original de: $\sqrt{0.23} = 0.479 \rightarrow e^{0.479} \approx 1.615$. En otras palabras, el precio predicho puede ser hasta un ~61.5% mayor o menor al original, desviación que tendrá mucha más relevancia cuanto mayor sea el rango de precio de los alojamientos. Este valor, sumado a un R-cuadrado de 0.24 para los valores predichos en la muestra de prueba, muestra que, aunque es útil para hacer aproximaciones al precio, sobre todo si este es bajo, aún existe un margen considerable de error.

Bibliografía

Anil Jadhav, Dhanya Pramod & Krishnan Ramanathan (2019) Comparison of Performance of Data Imputation Methods for Numeric Dataset, Applied Artificial Intelligence, 33:10, 913-933, DOI: 10.1080/08839514.2019.1637138

Gelman, A., & Hill, J. (2006). Missing-data imputation. En *Data Analysis Using Regression and Multilevel/Hierarchical Models* (pp. 529–544). Capítulo 25, Cambridge: Cambridge University Press.

Pham-Gia, T., & Hung, T. L. (2001). The mean and median absolute deviations. *Mathematical and computer Modelling*, *34*(7-8), 921-936.