Universidad de
SanAndrés

Big Data

Trabajo Práctico 2

Albareti Renato, López Federico y Mettola Franco

Fecha de entrega: 21 de abril de 2024

I. Limpieza

1)

a) Valores DuplicadosDuplicados

Eliminando duplicados, encontramos que habían 10 observaciones repetidas.

b) Columnas irrelevantes

Eliminiamos las variables `id`, `name`, `host_id`, `host_name`, `neighbourhood` y `last_review` ya que no aportan información relevante para el análisis.

c) Missing Values

Identificamos que para las variables `price` y `reviews_per_month` faltan datos. El problema de `reviews_per_month` era que para casas sin reseñas esta variable no estaba en 0, así que lo corregimos fácilmente. Pasamos al problema más serio de la variable `price`.

Como no tenemos información sobre el origen de esta base de datos, tendremos que especular sobre la causa de los valores faltantes que hubieran en la base de datos. Esto es importante porque si fue un problema en la descarga de los datos, es decir que por casualidad la computadora no imputó ciertos valores al azar, estamos en un caso de MCAR (missing completely at random). Si en cambio no se reportan por decisión de los reportantes, eso implica cierta sistematicidad que puede modelarse y solucionarse con métodos estadísticos, utilizando las variables para las cuales no faltan datos.

El punto importante es que podemos simplemente deshacernos de las observaciones (complete-case analysis) con valores faltantes en el primer caso, si son MCAR, porque de otra manera estaríamos introduciendo sesgo.

Nos decidimos por usar regresión linear ya que aunque es imposible probar aleatoreidad y tenemos una base de datos grande y varios predictores MAR puede ser un supuesto razonable. Consideramos es poco probable que los datos imputados sean extremadamente lejanos a los datos reales si los calculamos de esta manera. Lo implementamos con la clase de scikit-learn IterativeImputer usando regresion lineal iterativamente para obtener predicciones sobre las variables faltantes.

d) Outliers

No tiene sentido que un precio sea 0. Si alguno es 0 lo imputamos.

2) Métodos No Supervisados

a) Matriz de correlaciones

A continuación, presentamos la matriz de correlaciones. Con respecto a las variables categóricas, elegimos representarlas de manera individual en la matriz.

Con respecto a la imagen, observamos que el precio correlaciona positivamente con los alquileres de pisos enteros. Esto puede succeder ya que los huespedes prefieren alquilar la totalidad de los inmbuebles en lugar de habitaciones (ya sea compartida o privada). Por

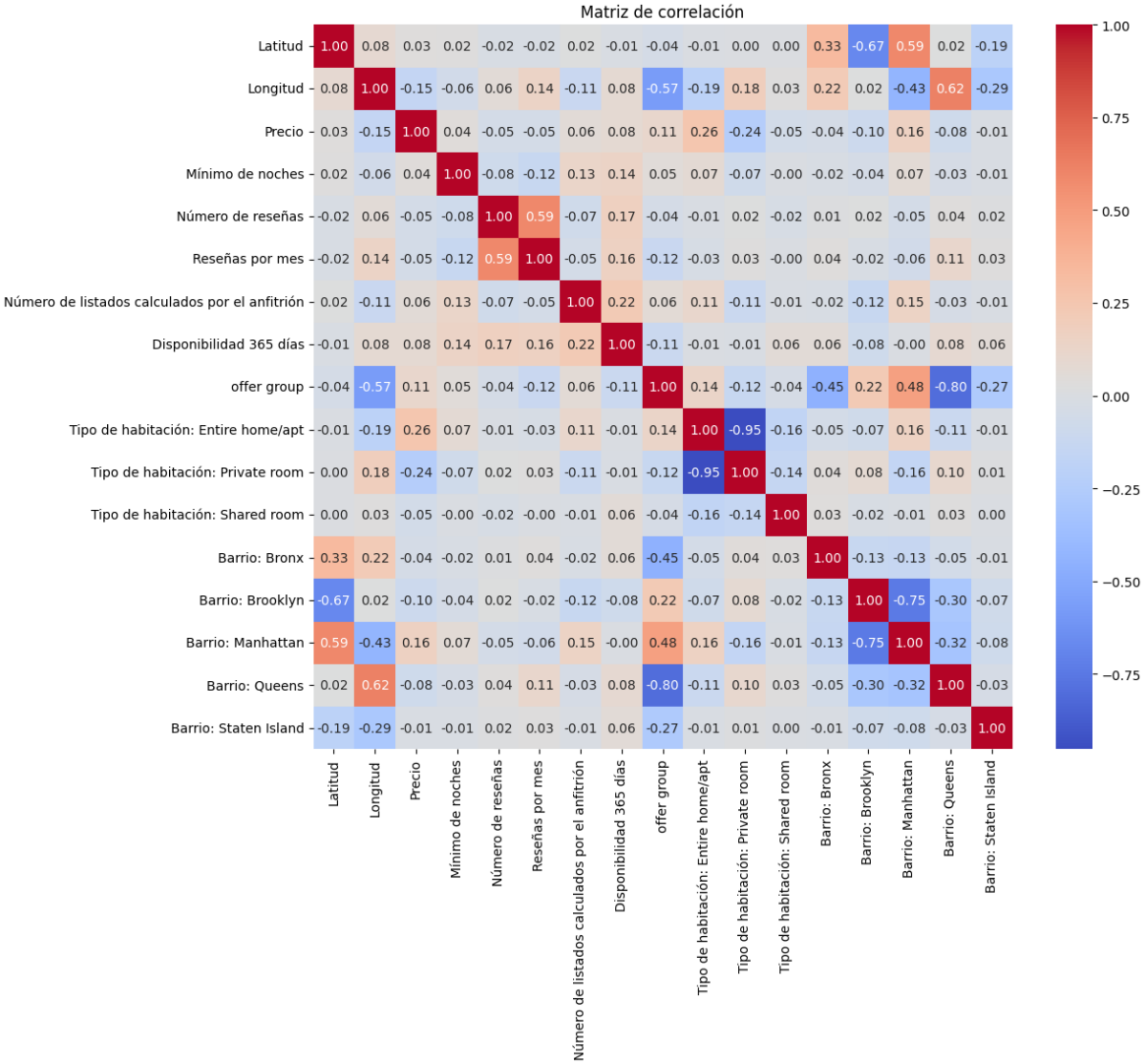

Figura 1: Matriz de Correlaciones

otro lado, también observamos una relación positiva con Manhattan, lo que tiene sentido dado que esta zona abarca la mayor parte de las secciones turísticas de Nueva York.

b) Oferentes por barrio y tipo de alquiler

Presentamos la siguiente tabla con la cantidad de departamentos por tipo de habitación. También se presenta un gráfico con la distribución.

Tipo de alquiler	Frecuencia
Piso Entero	25409
Habitación Privada	22326
Habitación Compartida	1160

Como se puede ver, las habitaciones privadas y pisos enteros conforman la gran mayoría

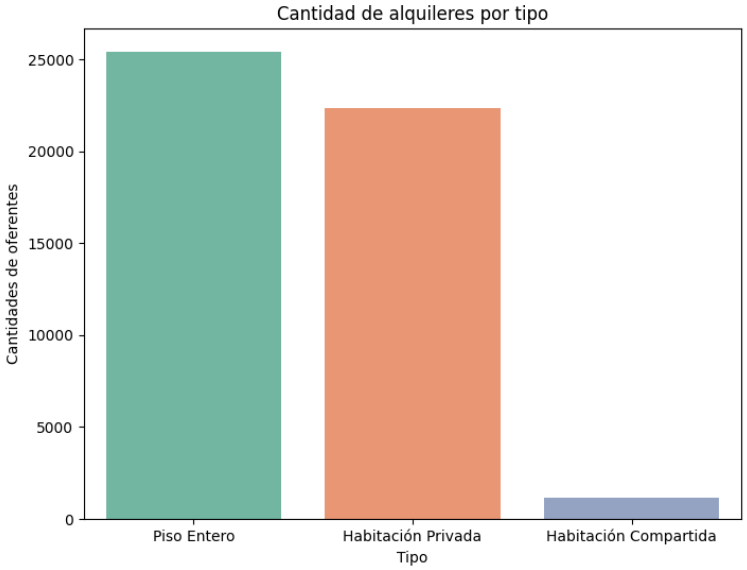

Figura 2: Cantidad de publicaciones por tipo

de ofertas en Airbnb en Nueva York. Esto nos informa que los usuarios priorizan la privacidad a la hora de reservar.

Con respecto a la distribución por barrio, presentamos la siguiente tabla y el siguiente gráfico.

Barrio	Frecuencia
Bronx	1091
Brooklyn	20104
Manhattan	21661
Queens	5666
Staten Island	373

En base a esto, destacamos que la gran mayoría de los alquileres están publicados en Manhattan y Brooklyn dado el atractivo turístico de estos barrios.

c) Histograma de precios

Presentamos el siguiente histograma donde se pueden ver la distribución de los precios en nuestra base de datos. En base a esto, observamos que los precios se acumulan en los valores menores a 1000 dólares. Cabe destacar que limitanos el rango del eje X para mejorar la visualización.

A continuación, presentamos los valores mínimos, máximos y la media de un alquiler por noche en Nueva York.

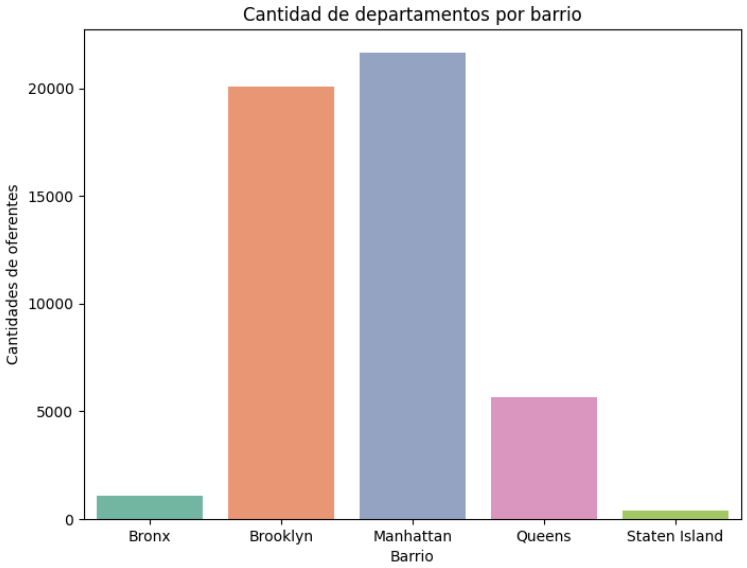

Figura 3: Cantidad de publicaciones por barrio

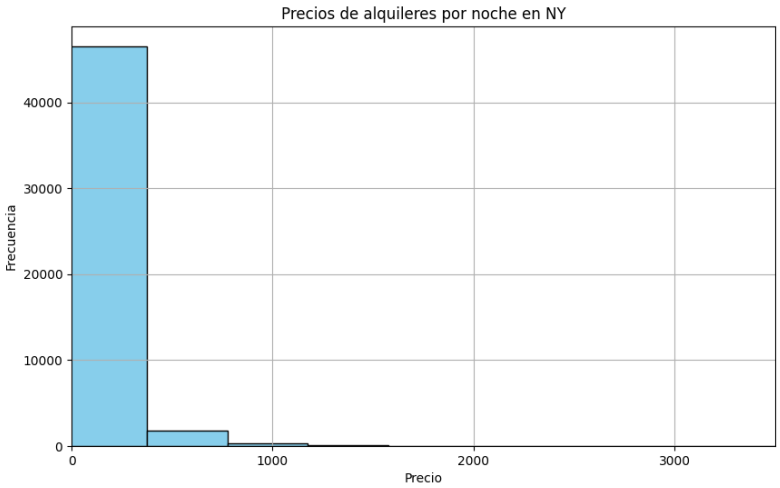

Figura 4: Distribución de los precios

Estadístico	Valor
Media	157.74
Precio Mínimo	10
Precio Máximo	10.000

Luego, presentamos las valores promedio de una noche en Nueva York condicional al tipo de inmueble.

Tipo de alquiler	Precio medio
Piso Entero	211,81
Habitación Privada	89.81
Habitación Compartida	70.11

Por último, presentamos el valor promedio de la reserva por cada barrio de Nueva York

Barrio	Precio medio
Bronx	87.60
Brooklyn	124.41
Manhattan	196.89
Queens	99.52
Staten Island	114.81

d) Scatter Plots

En el primer gráfico, presentamos la relación entre el precio del alquiler y la cantidad de reseñas recibidas. Cade destacar que estamos usando únicamente aquellas publicaciones con más de 100 reseñas recibidas. De esta manera, usamos a las publicaciones más revisadas de la plataforma. A modo de descripción, vemos que la mayor cantidad de puntos caen en valores cercanos a cero. Es decir, hay una gran cantidad de datos con valores bajos tanto de precio como de reseñas. Sin embargo, observamos una pendiente negativa que puede informar acerca de un problema de causalidad inversa. En otras palabras, los alquileres más baratos deberían atraer a más turistas y, por ende, tener más reseñas.

En este segundo scatter plot, relacionamos al precio del alquiler con la cantidad mínima de noches para reservar. En este caso, eliminamos del análisis a las publicaciones que no solicitan esta condición para alquilar. También limitamos el eje X a 3.000 dólares. Con respecto al gráfico, vemos una pendiente ligeramente negativa. A partir de esto, podemos intuir que los propietarios de los alquileres están dispuestos a reducir el precio con tal de asegurarse una estadía extensa.

e) PCA

A continuación, presentamos a los primeros dos componentes principales.

Con respecto al primer componente, se prioriza la variable longitud. Por otro lado, el segundo componente destaca a las variables relacionadas a las reseñas de la publicación.

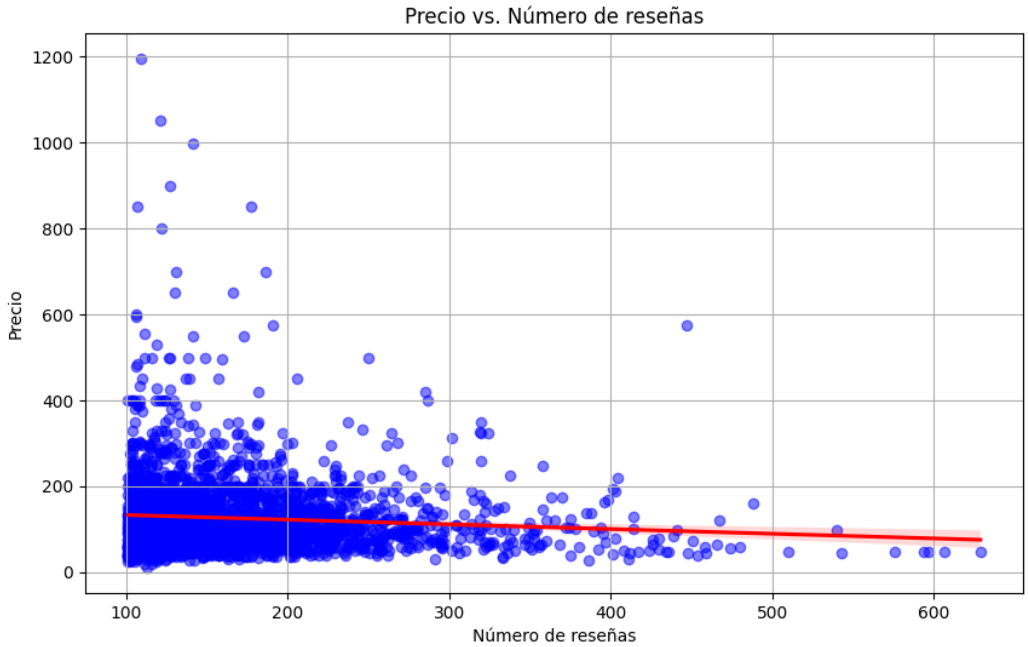

Figura 5: Scatter plot precio y reseñas

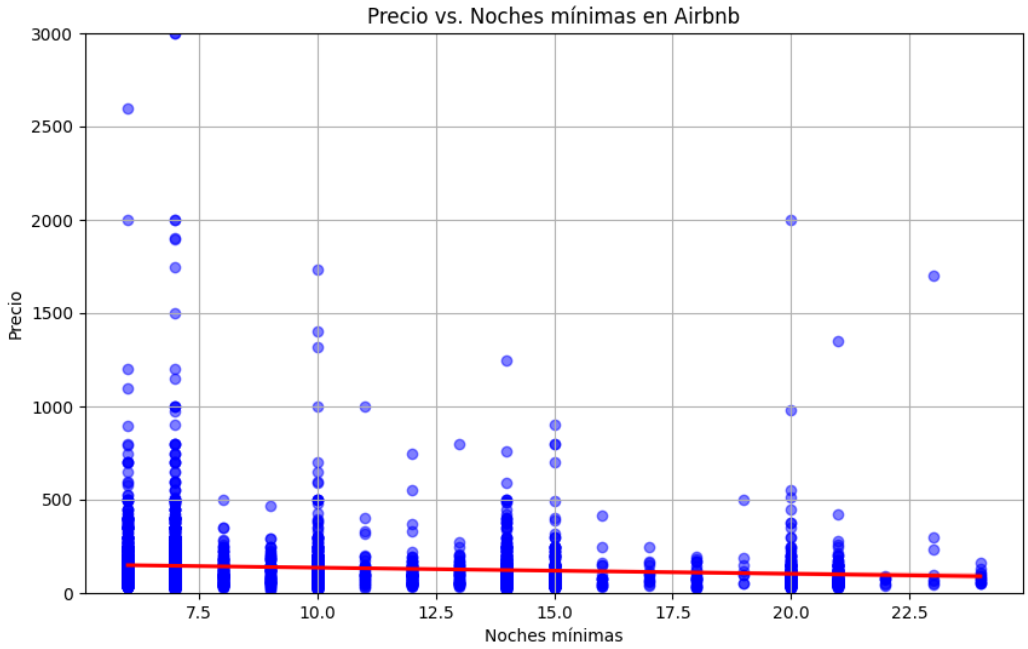

Figura 6: Scatter plot precio y reseñas

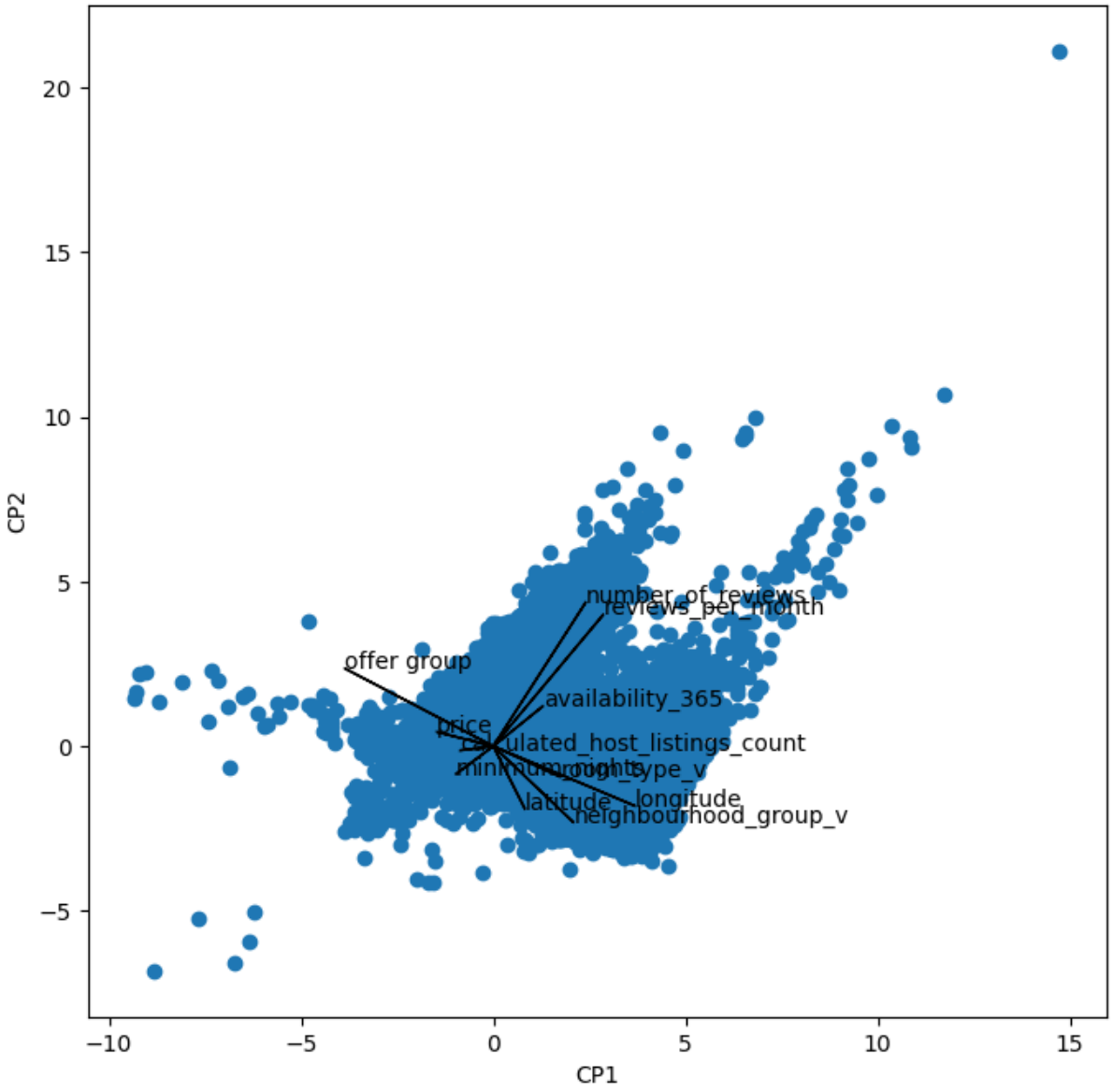

Figura 7: CP1 y CP2

Por lo tanto, las reseñas con altos puntajes del CP2 serán aquellas que más reseñas recibieron.

Con respecto a la varianza explicada, entre ambos suman 32 % de los cuales 18& pertenecen a CP1. Con respecto a los loadings, vemos que, como fue mencionado arriba, el primer componente le otorga una alta valoración a la variable longitud. Por otro lado, el segundo componente le brinda fundamental valoración a las reseñas recibidas. A partir del gráfico, vemos que los valores se concentran alrededor de los valores [0 , 5] en ambos componentes. Esto puede deberse a que los alquileres con bajo valor de longitud concentran un número escaso de reseñas recibidas. También observamos un grupo de publicaciones con altos valores en ambos componentes. Con respecto a los valores más negativos, vemos que son muy pocas las publicaciones ubicadas en valores negativos de CP1 y CP2.

3)

Dep. Variable:	price	**R-squared:**	0.092
Model:	OLS	**Adj. R-squared:**	0.092
Method:	Least Squares	**F-statistic:**	148.8
Date:	Sun, 21 Apr 2024	**Prob (F-statistic):**	1.53e-298
Time:	16:10:18	**Log-Likelihood:**	-1.0044e+05
No. Observations:	14669	**AIC:**	2.009e+05
Df Residuals:	14658	**BIC:**	2.010e+05
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P> |t|	[0.025	0.975]
const	-3.361e+04	4259.892	-7.891	0.000	-4.2e+04	-2.53e+04
Latitud	125.2279	36.495	3.431	0.001	53.694	196.762
Longitud	-386.8630	51.500	-7.512	0.000	-487.809	-285.917
Mínimo de noches	-0.0775	0.100	-0.772	0.440	-0.274	0.119
Número de reseñas	-0.2654	0.052	-5.135	0.000	-0.367	-0.164
Reseñas por mes	-2.1259	1.509	-1.408	0.159	-5.084	0.833
Número de listados calculados por el anfitrión	-0.1085	0.060	-1.800	0.072	-0.227	0.010
Disponibilidad 365 días	0.2210	0.015	14.632	0.000	0.191	0.251
Tipo de habitación	-99.1777	3.543	-27.992	0.000	-106.122	-92.233
Barrio	20.5085	2.964	6.920	0.000	14.699	26.318
Oferentes por barrio	0.0031	0.000	7.213	0.000	0.002	0.004

Omnibus:	32920.844	**Durbin-Watson:**	2.016
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	288282907.220
Skew:	21.036	**Prob(JB):**	0.00
Kurtosis:	688.486	**Cond. No.**	4.40e+07

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.4e+07. This might indicate that there are
strong multicollinearity or other numerical problems.

Observamos en la regresión que la mayoría de las variables son significativas, los casos que no cumplen significatividad son: Mínimo de noches con un P-valor ampliamente superior a cualquier banda de confianza, y Reseñas por mes, la cual también cuenta con un P-valor a aquel de un nivel de aceptación mínima (10 %).

Luego, la otra variable en tela de juicio es: Número de listados calculados por el anfitrión, que únicamente rechaza la hipótesis nula, de que su coeficiente estimado asociado

es realmente 0, a un nivel de confianza del 10 %.

El resto de las variables son significativas al 1 % en la definición del precio. Aún así, resaltamos que tanto el R^2 como el R^2 ajustado son iguales al 9,2 %, por lo que las variables utilizadas no logran representar un 10 % del precio publicado. Este hecho es acompañado por el coeficiente de curtosis, el cual muestra que las colas del modelo son pesadas, de manera que algún otro conjunto de factores que pueda explicar el precio mejor puede llegar a generar que el precio de un airbnb se encuentre a una distancia de la esperanza mayor a la varianza, es decir, en las colas de nuestro modelaje.