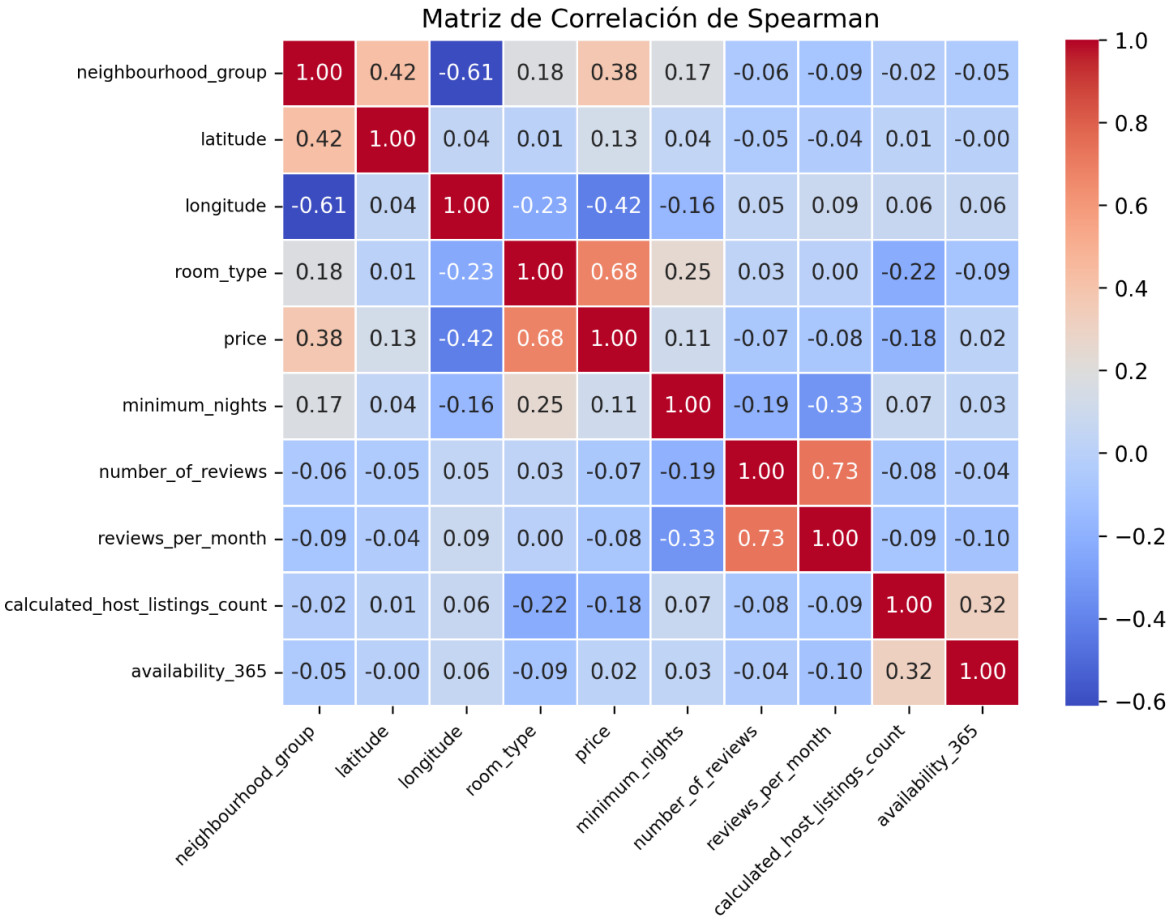
BIG DATA (UNT) -2024

TRABAJO PRÁCTICO N°2

Grupo 2: Fuentes Mortensen & Gorostiaga

Ejercicio 2

Gráfico 1: Matriz de correlación con el método de Spearman.



- Correlación entre reviews_per_month y number_of_reviews (0.73): se observa una correlación positiva y fuerte entre el número de reseñas por mes y el número total de reseñas, que responde a colinealidad entre estas dos variables que se da por su construcción.

- Correlación entre price y room_type (0.68): dado que la variable de tipo de habitación fue ordenada de peor a mejor tiene sentido que observemos una correlación positiva y fuerte entre el precio y esta variable categórica.

- Correlación entre price y neighbourhood_group (0.38): nuevamente, dado que la variable de vecindario fue ordenada de peor a mejor, también tiene sentido encontrar una correlación positiva, aunque menos fuerte que con room_type, entre el precio del alojamiento y el vecindario en el que está ubicado.

- Correlación entre reviews_per_month y minimum_nights (-0.33): la correlación entre la cantidad de reseñas por mes y cantidad mínima de noches es negativa. Esto podría dar cuenta de que a medida que aumenta el requisito de cantidad mínima de noches, el flujo de gente que ingresa al alojamiento por mes es menor y por ende también lo serán la cantidad de reseñas.

-Correlación entre room_type y mínimum nights (0.25): existe una correlación positiva aunque no tan fuerte entre estas variables. Aquellos alquileres que corresponden a peores alojamientos (menos privados) tienen menores requerimientos de días mínimos para poder reservar.

Ejercicio 3

Gráfico 2: Gráfico de barras sobre la proporción de oferentes de alojamientos en Airbnb, según vecindario.

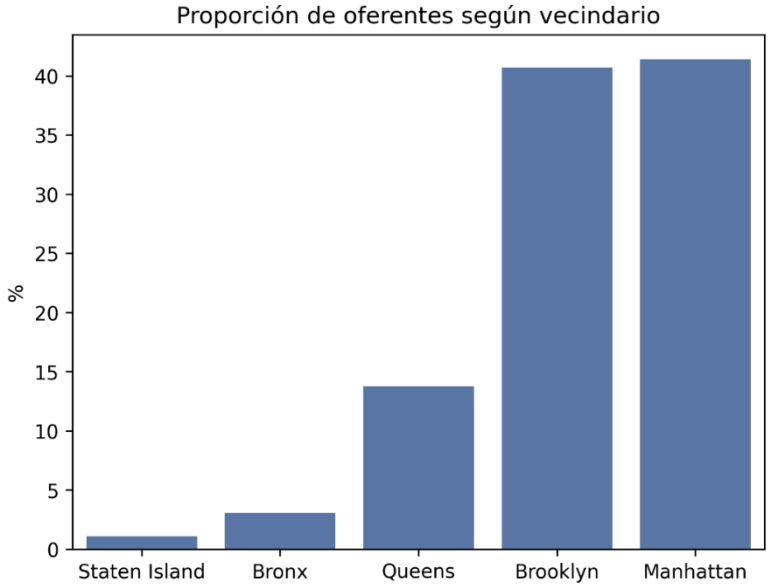
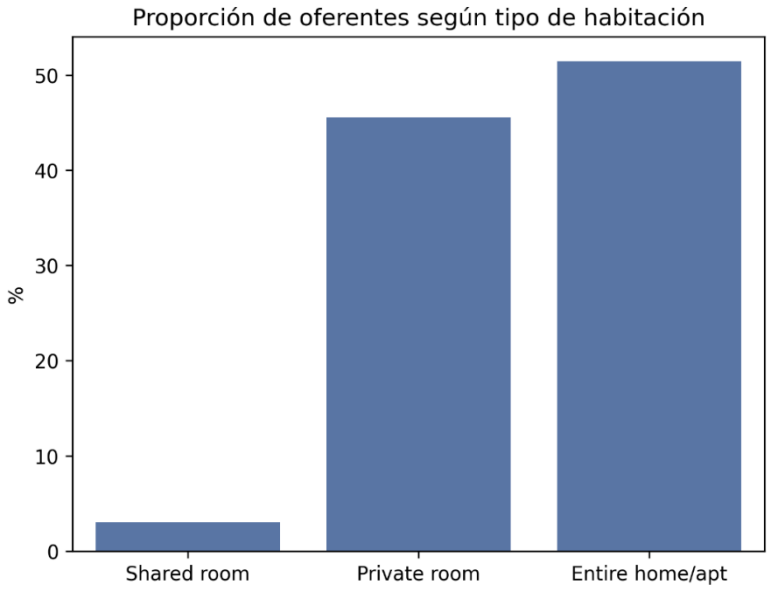


Gráfico 3: Gráfico de barras sobre la proporción de oferentes de alojamientos en Airbnb, según tipo de habitación.

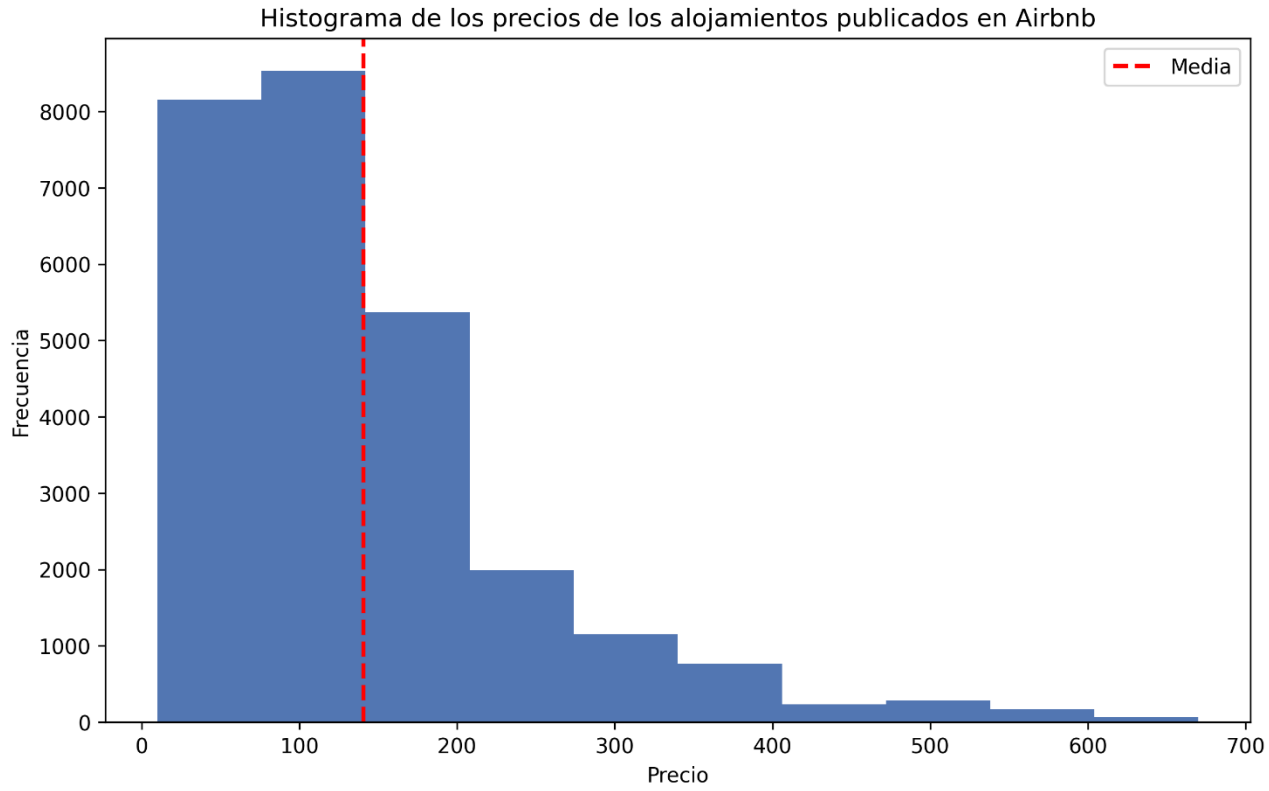


El gráfico 2 nos muestra que hay una mayor proporción de oferentes de alojamientos en Airbnb en Manhattan y Brooklyn en relación con Staten Island y Bronx.

En el gráfico 3 podemos ver que la proporción de oferentes que publica habitaciones compartidas es muy baja cuando se la compara con habitaciones privadas y casas o departamentos enteros.

Ejercicio 4

Gráfico 4: Histograma de los precios de los alojamientos publicados en Airbnb y su media.



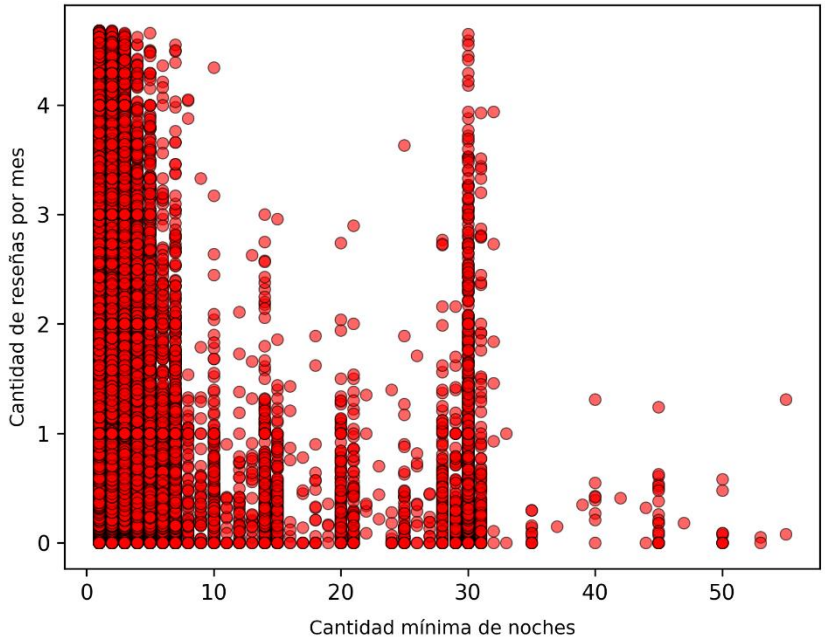
Los precios de los alojamientos publicados en Airbnb presentan una distribución asimétrica hacia la derecha, con precios que se concentran en valores menores a 200 dólares por noche. En la tabla 1 se presentan algunas estadísticas sobre la variable de precio, y su promedio según vecindario y según tipo de alojamiento.

Tabla 1: Estadísticas descriptivas del precio de los alojamientos de Airbnb para el total, según vecindario y según tipo de alojamiento.

PRECIO	Promedio
Total	140.63
Según vecindario:	
Bronx	85.51
Staten Island	93.02
Queens	98.02
Brooklyn	122.96
Manhattan	177.44
Según tipo de alojamiento:	
Habitación compartida	61.64
Habitación privada	86.02
Casa o departamento completo	193.57

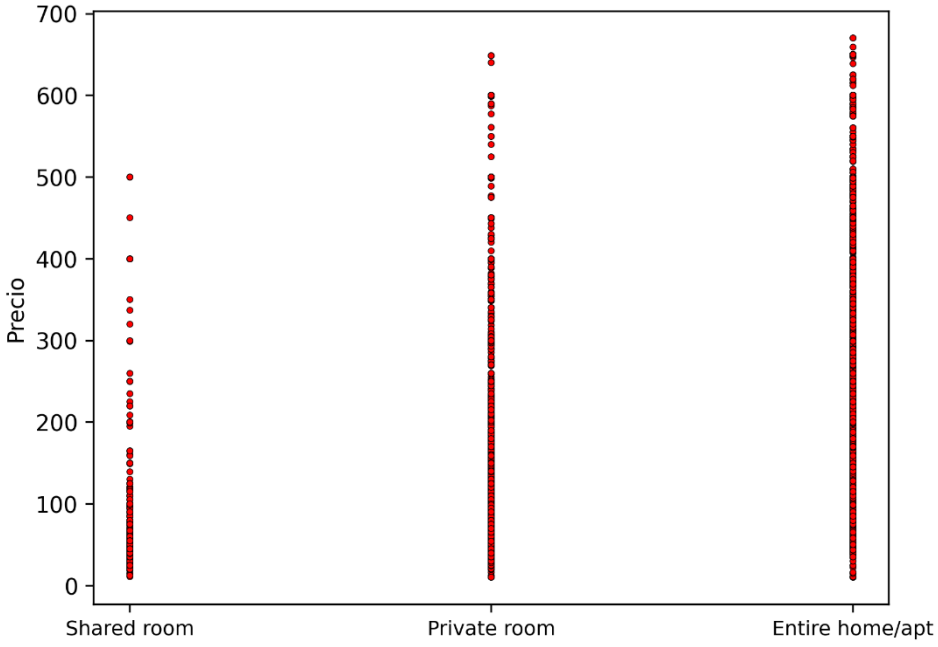
Ejercicio 5

Gráfico 5: Gráfico de dispersión entre la cantidad de reseñas por mes y la cantidad mínima de noches.



A pesar de la concentración de observaciones en 30 noches mínimas (propio de aquellos alquileres mensuales), se observa la correlación negativa aunque no muy fuerte entre cantidad mínima de noches y cantidad de reseñas por mes.

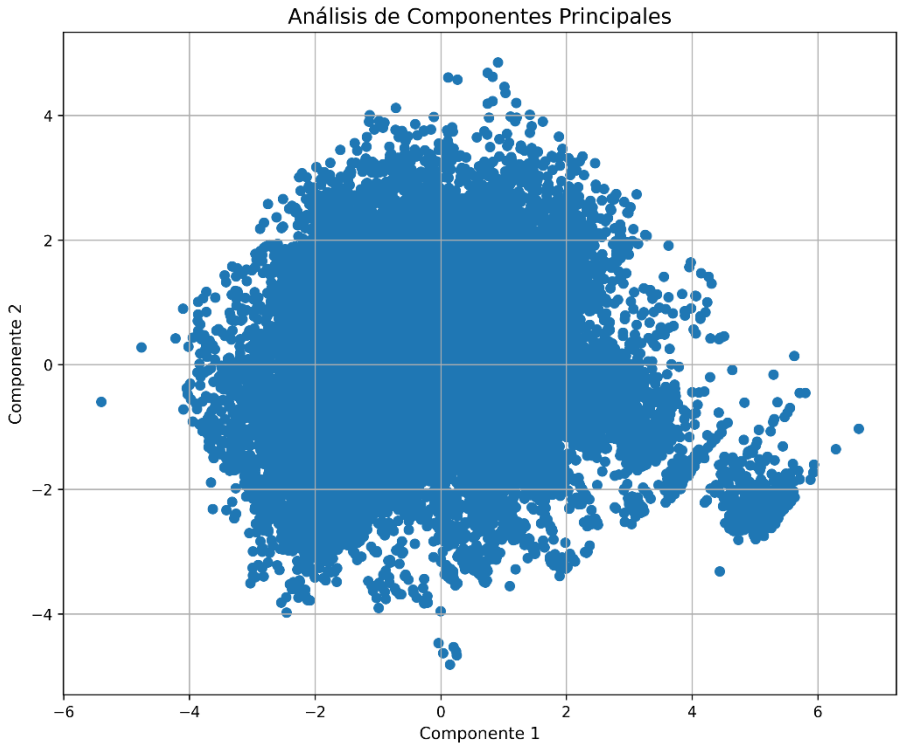
Gráfico 6: Gráfico de dispersión entre el precio de los alojamientos y el tipo de habitación.



Teniendo en cuenta que la variable de tipo de habitación es categórica, visualizamos una relación positiva entre el precio y el tipo de habitación con un primer salto más pronunciado entre habitación compartida y habitación privada. Sin embargo, la variedad de precios está presente en los tres tipos de alojamientos.

Ejercicio 6

Gráfico 7: Gráfico de dispersión entre el primer y segundo componente.



No identificamos ningún patrón específico al graficar ambos componentes. La nube de puntos observada sugiere variabilidad y complejidad en los datos.

El análisis de componentes principales (PCA) revela que el componente 1 explica el 22.62% de la varianza total de nuestro set de datos, mientras que el componente 2 explica un 17.54% adicional. En conjunto, los dos primeros componentes explican el 40.17% de la varianza total de los datos analizados.

Ejercicio 9

El modelo estimado, cuyos resultados se presentan en la tabla 2, presenta un coeficiente de determinación de aproximadamente 0.35. Esto indica que el 35.07% de la variabilidad en el precio de los alojamientos puede ser explicada por las variables independientes incluidas en el modelo.

Tabla 2: Resultados de la regresión lineal.

OLS Regression Results

Dep. Variable:	price	R-squared:	0.369
Model:	OLS	Adj. R-squared:	0.369
Method:	Least Squares	F-statistic:	1214.
Date:	Thu, 26 Sep 2024	Prob (F-statistic):	0.00
Time:	12:48:45	Log-Likelihood:	-1.0874e+05
No. Observations:	18699	AIC:	2.175e+05
Df Residuals:	18689	BIC:	2.176e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.36e+04	1193.663	-19.775	0.000	-2.59e+04	-2.13e+04
neighbourhood_group	18.5319	0.794	23.352	0.000	16.976	20.087
latitude	119.5567	10.708	11.165	0.000	98.568	140.546
longitude	-251.0947	13.956	-17.991	0.000	-278.450	-223.739
room_type	94.2382	1.102	85.489	0.000	92.077	96.399
minimum_nights	-1.7449	0.074	-23.660	0.000	-1.889	-1.600
number_of_reviews	-0.2853	0.024	-12.111	0.000	-0.331	-0.239
reviews_per_month	-3.3802	0.612	-5.527	0.000	-4.579	-2.181
calculated_host_listings_count	-0.1051	0.054	-1.942	0.052	-0.211	0.001
availability_365	0.1145	0.005	23.332	0.000	0.105	0.124

Omnibus:	8188.678	Durbin-Watson:	2.004
Prob(Omnibus):	0.000	Jarque-Bera (JB):	45159.811
Skew:	2.069	Prob(JB):	0.00
Kurtosis:	9.391	Cond. No.	4.55e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.55e+05. This might indicate that there are
strong multicollinearity or other numerical problems.