

Trabajo práctico Integrador

**Materia “Introducción al Aprendizaje Estadístico”** *Diplomatura en Ciencias Sociales Computacionales.* Universidad Nacional Guillermo Brown - UnaB.

Integrantes:

Florencia Palacio

Claudio Meschini

Marianela Pi

Contenido

[Objetivo. 3](#_Toc140685001)

[Consignas. 3](#_Toc140685002)

[PARTE I. Estadística descriptiva. 3](#_Toc140685003)

[¿Qué columnas tiene?¿Qué tipo de columnas son? 3](#_Toc140685004)

[a) Análisis y aplicación de Medidas de Tendencia Central. 3](#_Toc140685005)

[*b)* Análisis y aplicación de *Medidas de Dispersión.* 5](#_Toc140685006)

[c) Análisis de covarianza entre las variables. 7](#_Toc140685007)

[PARTE II. Modelado supervisado. 9](#_Toc140685008)

[*a.* Modelo que prediga el *precio.* 9](#_Toc140685009)

[*b.* Modelo que prediga el *barrio.* 11](#_Toc140685010)

[PARTE III. Modelado no supervisado. 13](#_Toc140685011)

[a. Implementar el modelo de acuerdo a las variables de agrupamiento o predictoras que consideren mejor. 13](#_Toc140685012)

[b. Evaluar cuál es el mejor número de clusters para agrupar el conjunto de datos. 14](#_Toc140685013)

# Objetivo.

El presente trabajo tiene como finalidad la integración y aplicación de las técnicas dictadas durante el Modulo de Introducción al Aprendizaje Estadístico como así también la aplicación correcta de los algoritmos.

# Consignas.

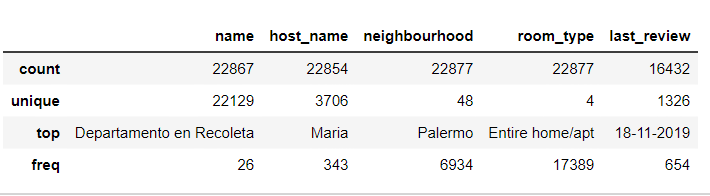
El trabajo será un informe estadístico de 3 segmentos que contenga la respuesta a los siguientes conceptos. Van analizar los datos del conjunto de datos “**Buenos Aires Airbnb Open Data”.** [**(Link al Kaggle)**](https://www.kaggle.com/datasets/sameerkulkarni91/buenos-aires-airbnb-data)

# PARTE I. Estadística descriptiva.

## ¿Qué columnas tiene?¿Qué tipo de columnas son?

Tenemos 10 variables numéricas (3 son flotantes y 7 enteras) y 5 categóricas

### Análisis y aplicación de Medidas de Tendencia Central.



Estos datos proporcionan un resumen de las frecuencias y las categorías más comunes en cada una de las columnas categóricas del conjunto de datos.

Variable neighbourhood: Hay un total de 22.877 registros en la columna 'neighbourhood'. Hay 48 valores únicos en esta columna, lo que indica que hay 48 barrios diferentes. El barrio más común es "Palermo", que aparece 6,934 veces.

Variable room\_type: Hay un total de 22.877 registros en la columna 'room\_type'. Hay 4 valores únicos en esta columna, lo que indica que hay 4 tipos de habitaciones diferentes. El tipo de habitación más común es "Entire home/apt" (departamento completo), que aparece 17,389 veces.

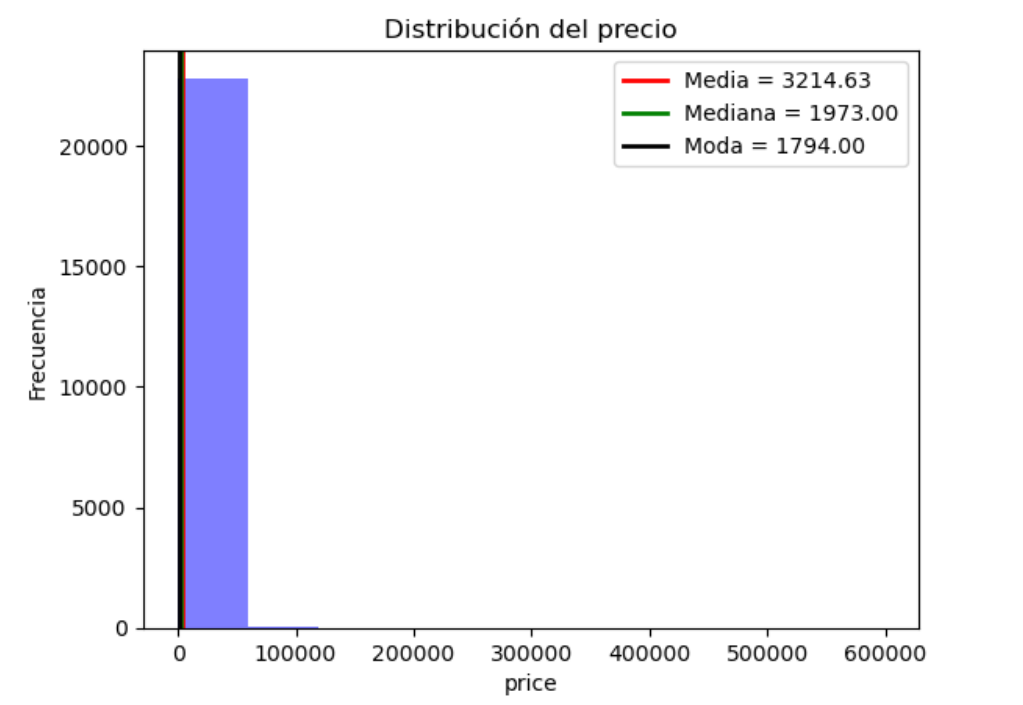
En la siguiente tabla



Los resultados muestran que el precio promedio de los alojamientos es de aproximadamente 3.214,63 unidades monetarias, con un rango que va desde 120 hasta 597.865 unidades monetarias. En cuanto a la duración mínima de noches, se observa un valor promedio de aproximadamente 5,65 noches, variando desde 1 noche hasta 1.125 noches.

En cuanto a las revisiones de los alojamientos, se encontró un promedio de 15.48 revisiones por mes, con un rango que va desde 0 hasta 600 revisiones. Esto indica un nivel significativo de participación y retroalimentación por parte de los huéspedes. Además, el promedio de anuncios calculados por los anfitriones es de aproximadamente 6,30, con un mínimo de 1 y un máximo de 105 anuncios.

En cuanto a la disponibilidad de los alojamientos a lo largo del año, se observa un promedio de aproximadamente 202,06 días disponibles, con un rango que va desde 0 hasta 365 días. Esto indica una oferta relativamente amplia de alojamientos en la plataforma.



Podemos observar que la media es de 3214.63, lo que indica un valor promedio relativamente alto. La moda es de 1794 lo que sugiere que este valor es el más frecuente en la distribución. La mediana, que es de 1973.0, indica que la mitad de los precios se encuentran por encima de este valor y la otra mitad por debajo de él.

En el histograma, podemos apreciar una curtosis asimétrica hacia la derecha, lo que indica que hay una concentración de valores más bajos en la cola derecha de la distribución. La mayoría de los registros muestran precios que se agrupan alrededor del rango de 0 a 50000, lo que indica una tendencia a concentrarse en esa franja de precios.

La cantidad mínima de precio es: 120

La cantidad máxima de precio es: 597865

El rango de precio es: 597745

El error standard de precio es: 12364.1596

Calculamos la varianza de todas las variables numéricas

id 1.418935e+14

host\_id 8.600790e+15

latitude 3.350545e-04

longitude 8.962125e-04

price 1.528791e+08

minimum\_nights 4.628168e+02

number\_of\_reviews 9.777231e+02

reviews\_per\_month 2.074837e+00

calculated\_host\_listings\_count 2.175330e+02

availability\_365 1.799936e+04

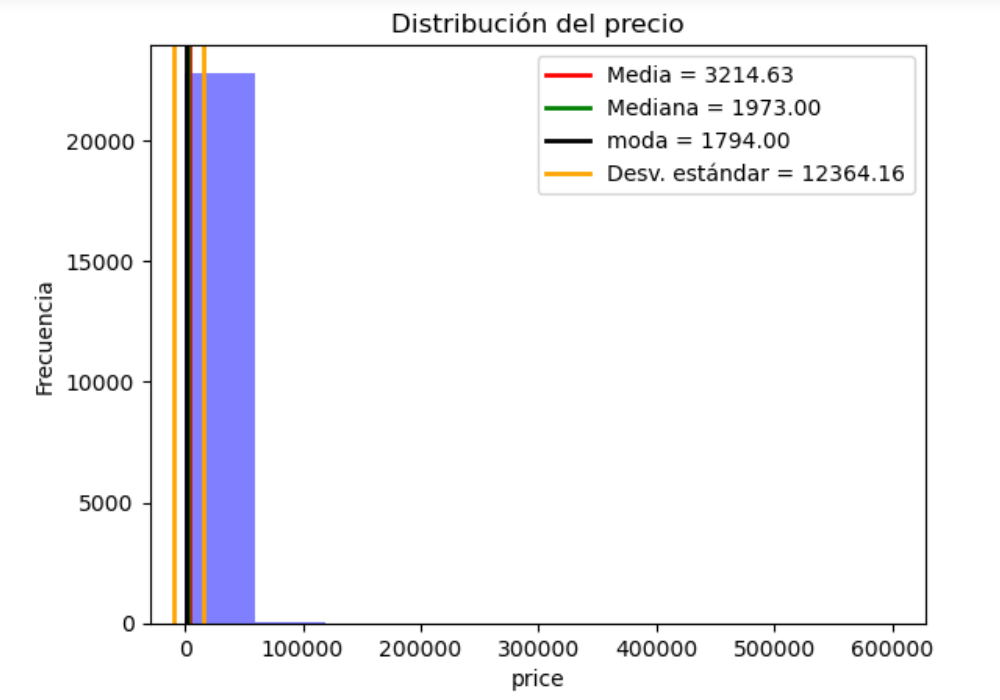
Curtosis

1442.6202036326044

Un valor positivo de curtosis, como el que se ha calculado (1442.6202036326044), indica una distribución más concentrada y puntiaguda en comparación con la

distribución normal.

### Análisis y aplicación de *Medidas de Dispersión.*



Desviación estándar: 12364.15963950588

Varianza: 152872443.59118617

Rango intercuartil: 1794.0

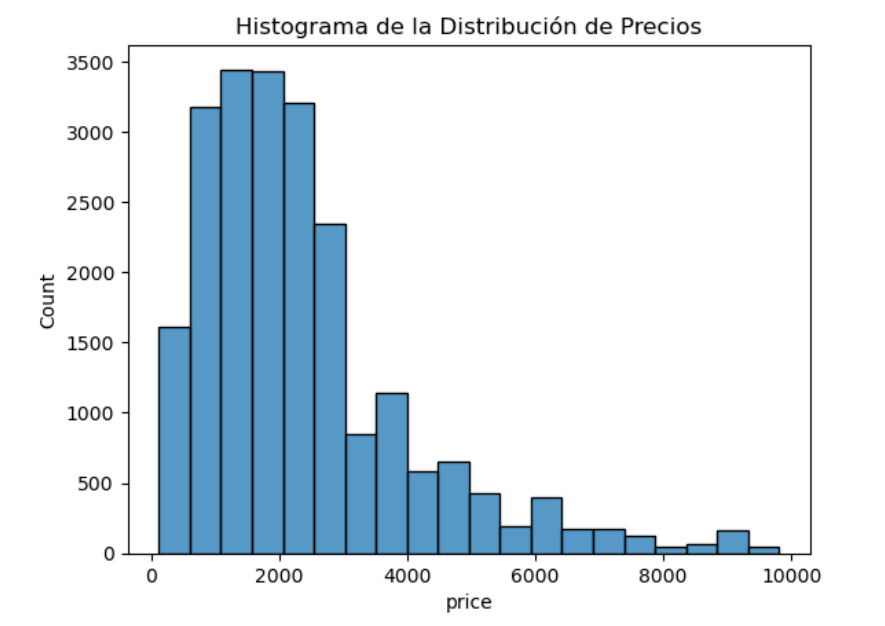
Coeficiente de variación: 384.621523616117

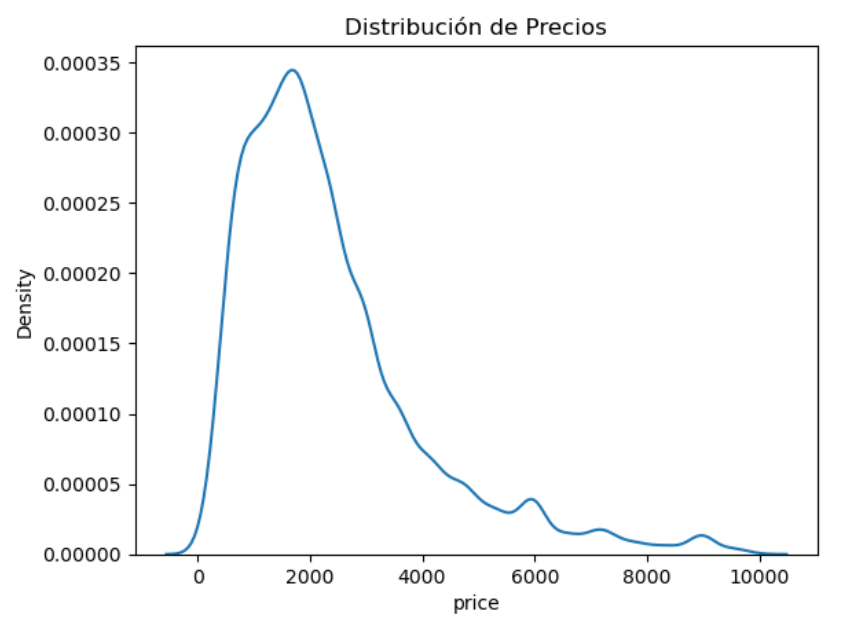
La desviación estándar de 12364.1596 nos indica que los precios tienen una dispersión considerable en relación con la media. La varianza de 152872443.5912 también refuerza esta idea.

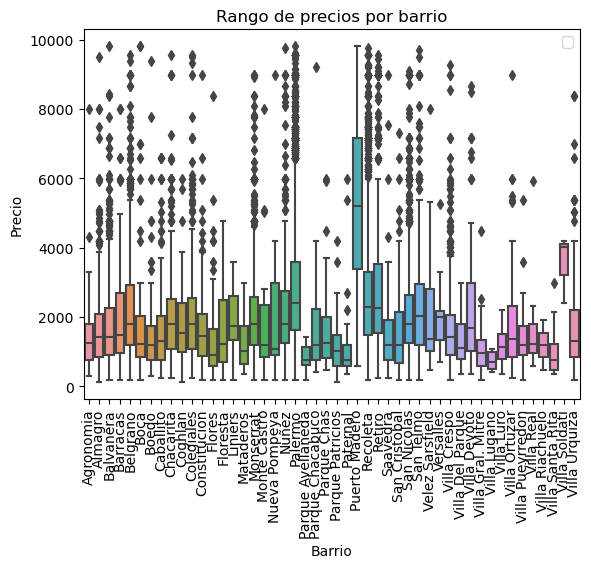
El rango intercuartil, que es de 1794.0, indica que la mitad de los precios se encuentra dentro de este rango relativamente estrecho.

El coeficiente de variación de 384.6215 nos indica que la variabilidad relativa de los precios es alta en relación con la media.

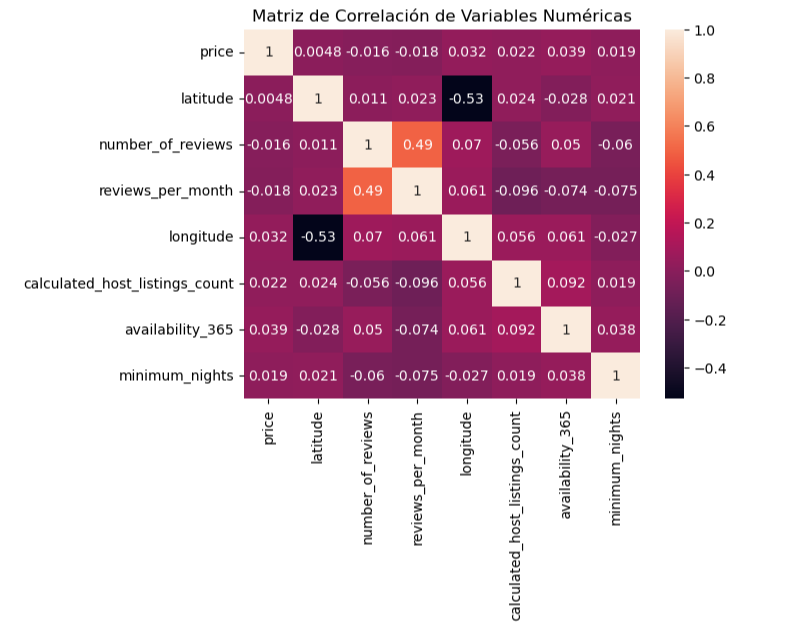
Si consideramos los casos cuyo precio es hasta 5 veces la mediana, estamos incluyendo el 98% de los casos. Esto nos permite observar la dispersión de los precios con más detalle y tener una idea más precisa de la variabilidad en los datos.







### Análisis de covarianza entre las variables.



* price 1.000000
* minimum\_nights 0.019407
* number\_of\_reviews -0.015942
* reviews\_per\_month -0.018156
* calculated\_host\_listings\_count 0.022110
* availability\_365 0.038567

La variable "minimum\_nights" tiene una correlación positiva muy baja con el precio (0.019407). Esto indica que un aumento en el número mínimo de noches generalmente tiene una influencia mínima en el precio.

La variable "number\_of\_reviews" tiene una correlación negativa muy baja con el precio (-0.015942). Esto sugiere que la cantidad de reseñas no tiene una relación fuerte con el precio.

La variable "reviews\_per\_month" también tiene una correlación negativa muy baja con el precio (-0.018156). Esto implica que la tasa de reseñas por mes tiene una influencia mínima en el precio.

La variable "calculated\_host\_listings\_count" muestra una correlación positiva baja con el precio (0.022110). Esto sugiere que un mayor número de anuncios calculados por el host puede tener una ligera influencia en el precio.

La variable "availability\_365" tiene una correlación positiva baja con el precio (0.038567). Esto indica que una mayor disponibilidad durante todo el año podría tener una influencia ligeramente positiva en el precio.

En general, basándonos en estos coeficientes de correlación, ninguna de las variables tiene una relación fuerte con el precio. Sin embargo, las variables "calculated\_host\_listings\_count" y "availability\_365" parecen tener una correlación ligeramente más alta en comparación con las demás.

Calculamos la covarianza entre la variable availability\_365 y el precio

La covarianza es: 63975.397716320316

La covarianza de 63975.39 es positiva y nos indica que hay una relación directa entre estas dos variables, lo cual significa que a medida que el mínimo de noches aumenta, también tiende a aumentar el precio. Sin embargo, hay que tener en cuenta que la covarianza no proporciona una medida normalizada de la fuerza de la relación entre las variables.

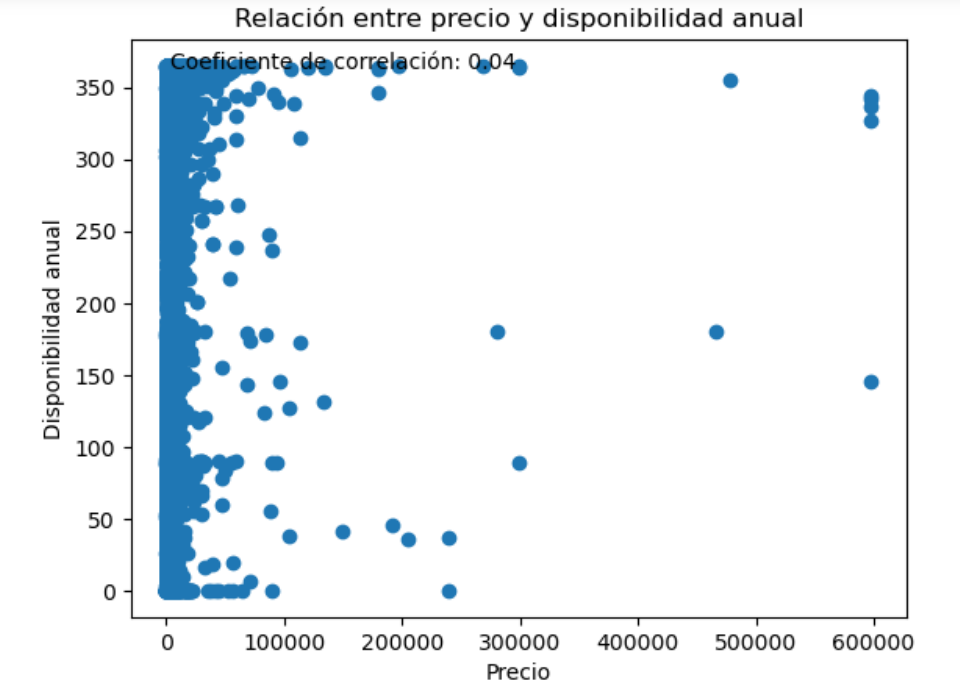
Calculamos la correlación entre la disponibilidad anual y el precio

La matriz de correlación entre la disponibilidad anual y el precio es:

[[1. 0.03856652]

[0.03856652 1. ]]

La matriz de correlación es simétrica, por lo que el valor en la posición (0, 1) es el mismo que en la posición (1, 0), en este caso es aproximadamente 0.038566517663953435. Este valor indica una correlación muy débil y cercana a cero entre las dos variables.



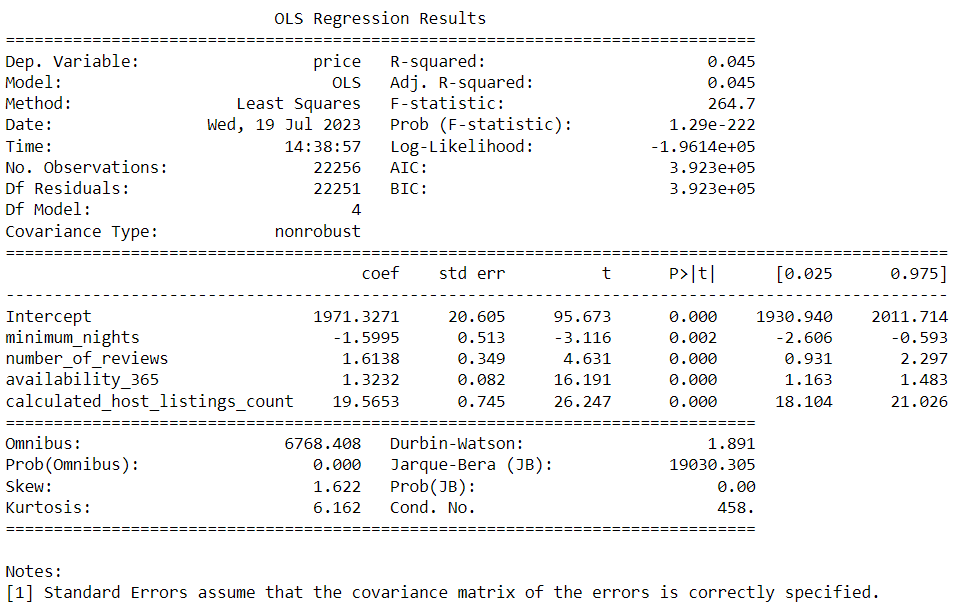
# PARTE II. Modelado supervisado.

Se solicita generar dos modelos de aprendizaje supervisado.

### Modelo que prediga el *precio.*

Creación y entrenamiento del modelo de predicción de precios:

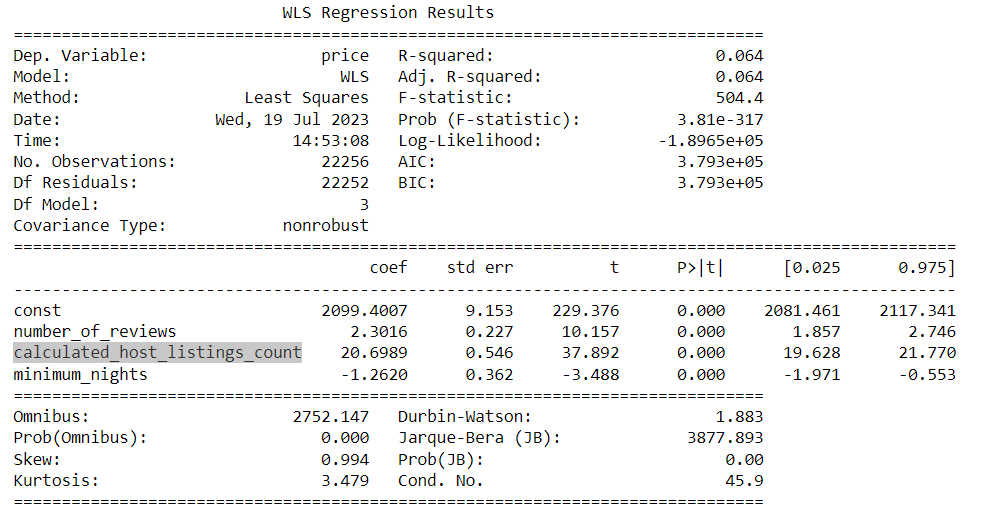
Tomamos las variables minimum\_nights + number\_of\_reviews + availability\_365 y calculated\_host\_listings\_count para hacer la regresión OLS



Podemos observar que el valor p de mínimos de noche no es igual a 0.000 y por eso hacemos una regresión ponderada solo con las variables minimum\_nights + number\_of\_reviews + calculated\_host\_listings\_count.

Realizamos un ajuste de regresión lineal ponderada (weighted linear regression)

Este código permite obtener un ajuste de regresión lineal ponderada, donde se asigna mayor importancia a las observaciones con residuos más pequeños. De esta manera logramos subir de 0.045 a 0.065

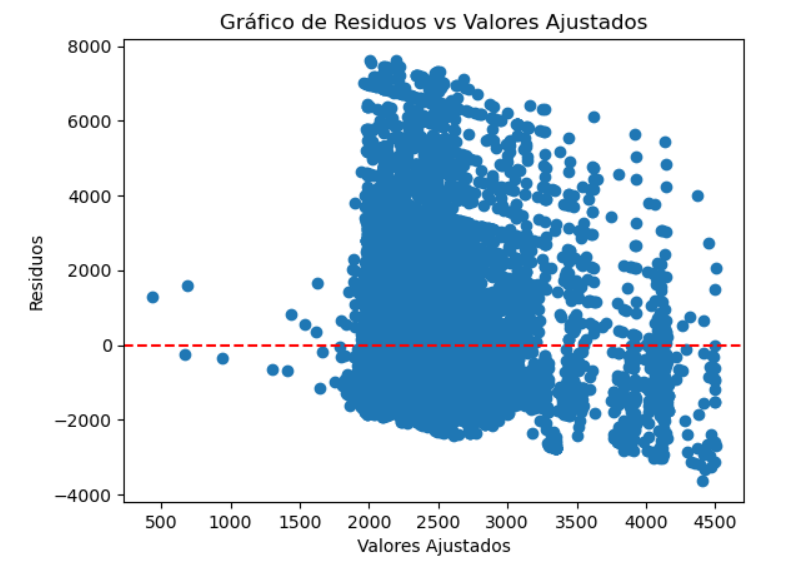


En este caso, el R-cuadrado es 0.065, lo que indica que aproximadamente el 6.5% de la variabilidad en el precio puede ser explicada por las variables minimum\_nights, number\_of\_reviews y calculated\_host\_listings\_count. En otras palabras, el modelo de regresión ponderada con las variables minimum\_nights, number\_of\_reviews y calculated\_host\_listings\_count pueden explicar solo una pequeña proporción de la variabilidad en el precio.

Estadístico de la prueba: 128.32261601209305

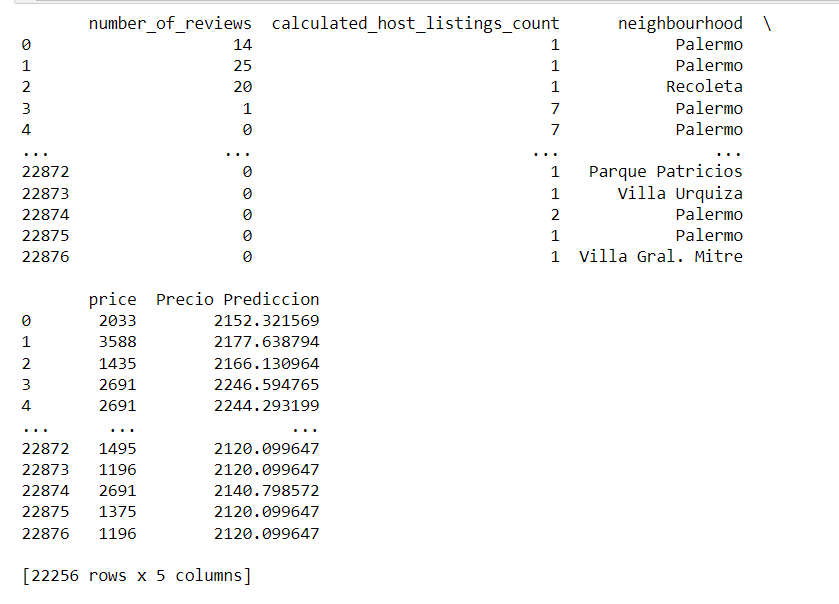
Valor p de la prueba: 8.893808514189698e-27

Valor crítico de la prueba: 7.49253352600262e-27



En este caso, el valor p de la prueba (8.893808514189698e-27) es menor que el nivel de significancia comúnmente utilizado (0.05), lo que sugiere que hay evidencia suficiente para rechazar la hipótesis nula y afirmar que existe una relación significativa entre las variables en el modelo de regresión ponderada.

Hicimos el cálculo con las diferentes variables, pero el precio que mas se acercaba era calculándolo con la variable numero de reviews y calculated\_host\_listings\_count



El error de la prueba: 0.7079713680522782

El resultado obtenido es 0.7079713680522782. Este valor representa el cociente entre el error estándar de la estimación y el promedio de los valores reales. Un error más bajo indica que las predicciones del modelo tienen menos dispersión con respecto a los valores reales, lo que implica un mejor ajuste del modelo.

Modelo para predecir el precio:

MSE: 2587923.9020174453

R²: 0.036620465873922425

Según los resultados, el modelo tiene un error cuadrático medio (MSE) de aproximadamente 2587923.9020174453 y un coeficiente de determinación (R²) de aproximadamente 0.0044.

Esto indica que el modelo no se ajusta bien a los datos y no puede explicar una gran parte de la variabilidad en los precios. El MSE alto sugiere que las predicciones del modelo tienen una gran discrepancia con los valores reales de los precios.

### Modelo que prediga el *barrio.*

#### Seleccionar variables y aplicar modelos de aprendizaje supervisado de acuerdo a la combinación de variables encontradas.

Aquí están las variables utilizadas:

<class 'pandas.core.frame.DataFrame'>

Int64Index: 22256 entries, 0 to 22876

Data columns (total 4 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 price 22256 non-null int64

1 id\_room\_type 22256 non-null category

2 rpm\_sin\_nan 22256 non-null float64

3 minimum\_nights 22256 non-null int64

En cuanto a la selección de variables y la elección del modelo de aprendizaje supervisado, se ha utilizado un conjunto de cuatro variables predictoras: "price", "id\_room\_type", "rpm\_sin\_nan" y "minimum\_nights". Estas variables han sido preparadas y utilizadas para entrenar un modelo KNN (KNeighborsClassifier).

Luego, dividimos los datos en conjuntos de entrenamiento y prueba utilizando un método random con una proporción del 75% para el entrenamiento y el 25% para la prueba.

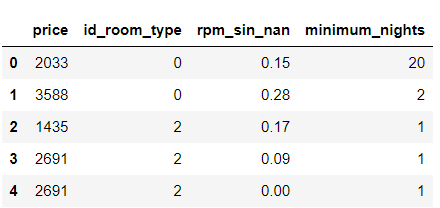
También calculamos métricas de evaluación, como precisión, recall, f1-score y la matriz de confusión, para comprender la calidad de las predicciones y cómo el modelo se desempeña en cada barrio.

Finalmente, utilizamos el modelo KNN para entrenar y predecir los barrios en el conjunto de prueba.

#### Incluir las métricas de evaluación y fundamentos de por qué se seleccionaron las variables predictoras.

El barrio que se predijo el modelo fue Belgrano

array(['Belgrano'], dtype=object)



La precisión del modelo es de 0.20619946091644206, significa que el modelo KNN que has entrenado tiene un rendimiento bajo en la predicción de los datos de prueba.

En el reporte de clasificación que se encuentra disponible en el Notebook , puedes observar las siguientes métricas para cada barrio:

Precisión: Es la proporción de predicciones positivas correctas (verdaderos positivos) sobre todas las predicciones positivas (verdaderos positivos + falsos positivos).

Recall: Es la proporción de predicciones positivas correctas (verdaderos positivos) sobre todas las muestras que realmente son positivas (verdaderos positivos + falsos negativos).

F1-score: Es la media armónica entre la precision y el recall, y proporciona una medida del equilibrio entre ambas métricas.

Support: Es el número de muestras reales de cada clase (barrio) en el conjunto de prueba.

Además, la matriz de confusión muestra el número de predicciones para cada combinación de clases (barrios) en el conjunto de prueba.

Teniendo en cuenta que el accuracy obtenido es del 21%, podemos concluir que el modelo KNN no está obteniendo buenos resultados en la predicción de los barrios. Es probable que sea necesario ajustar los hiperparámetros del modelo o probar con otros algoritmos de clasificación para mejorar el rendimiento.

Tipo de habitación: La variable "room\_type" indica si el alojamiento es un apartamento completo, una habitación privada, una habitación compartida, etc. Esta variable puede ser relevante, ya que los diferentes barrios pueden tener una distribución característica de los tipos de habitaciones.

Precio: La variable "price" puede ser una variable importante para predecir el barrio, ya que los precios de los alojamientos pueden variar según la ubicación. Es posible que ciertos barrios tengan un rango de precios más alto o más bajo, lo que podría ser útil para distinguir entre ellos.

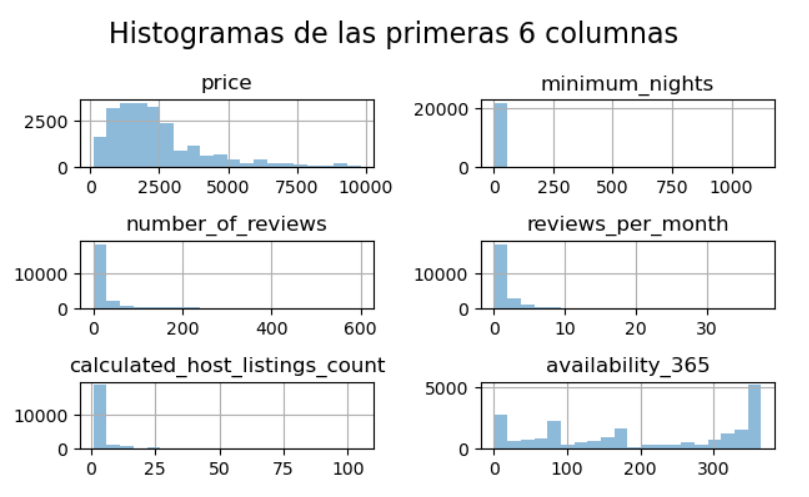
Número de reseñas y revisión mensual: Las variables "number\_of\_reviews" y "reviews\_per\_month" pueden proporcionar información sobre la popularidad y la demanda de los alojamientos en diferentes barrios. Pueden indicar la cantidad de visitantes y la satisfacción de los huéspedes, lo cual puede estar relacionado con las características únicas de cada barrio.

## PARTE III. Modelado no supervisado.

Se solicita generar un modelo de ***agrupamiento.***

### Implementar el modelo de acuerdo a las variables de agrupamiento o predictoras que consideren mejor.

El modelo que vamos a usar es KMeans : El modelo KMeans es un algoritmo de agrupamiento no supervisado que busca particionar los datos en k grupos o clusters. Al crear una instancia de KMeans, puedes ajustar diferentes parámetros para controlar cómo se llevará a cabo el proceso de agrupamiento. Aquí están los parámetros más importantes que se pueden ajustar en KMeans:



### Evaluar cuál es el mejor número de clusters para agrupar el conjunto de datos.

El gráfico del método del codo, busca el punto donde la inercia (suma de las distancias) deja de disminuir rápidamente y forma un "codo". En el gráfico del coeficiente de silueta, busca el valor máximo del coeficiente.

El mejor número de clusters según el gráfico es 4.

