

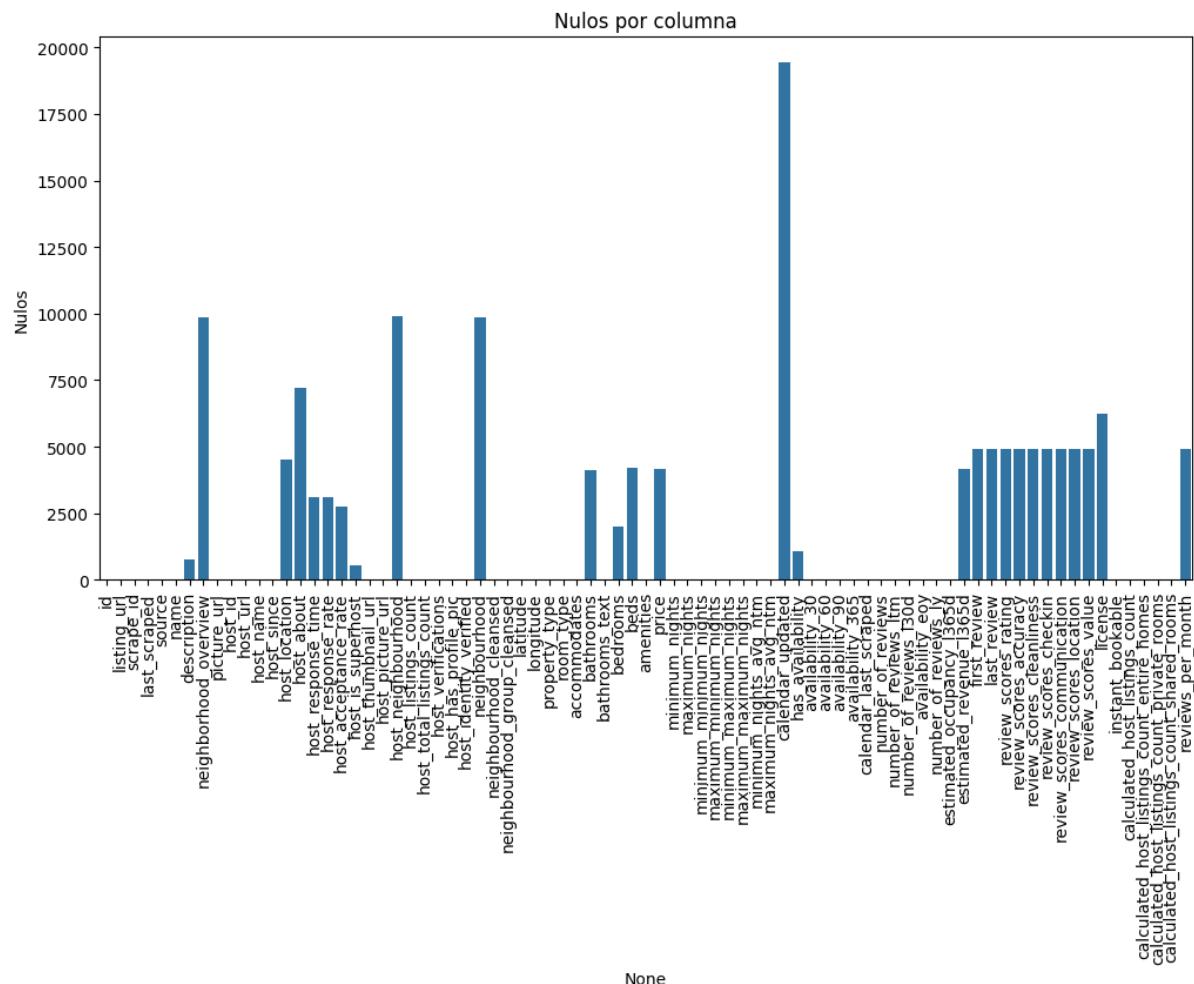
# 🔍 Análisis Exploratorio de Airbnb Barcelona (Exploratory Data Analysis - EDA)

Este análisis exploratorio tiene como objetivo comprender mejor los datos de Airbnb en Barcelona, identificando patrones, tendencias y áreas de interés a través de visualizaciones y estadísticas descriptivas.

## 💻 Archivos Analizados

Listings.csv: 19.422 registros, 79 columnas.

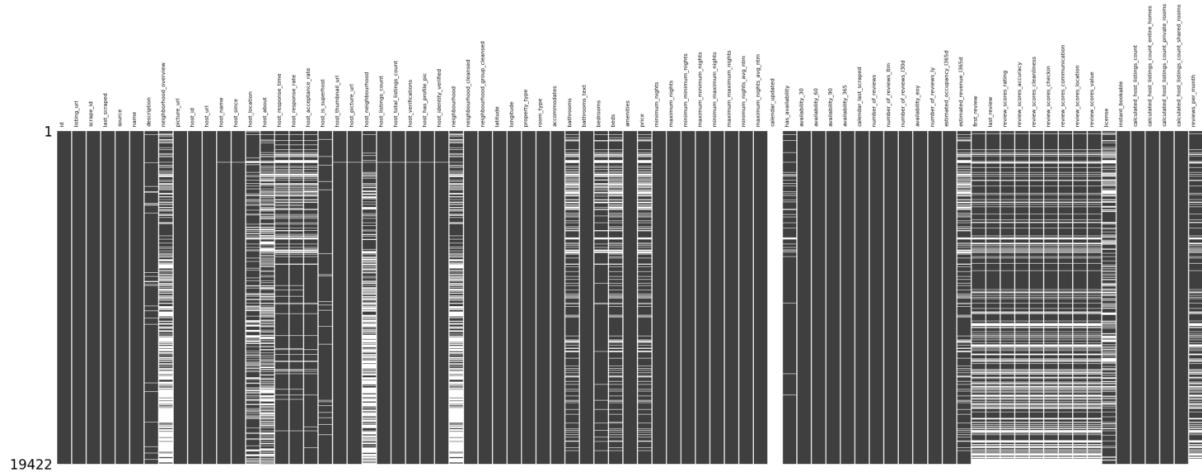
## ⚠️ Observacion de cantidad nulos



A continuacion se muestran los nulos encontrados en el dataset Listings.csv usando un grafico de barras que nos ayuda a identificar las columnas con mayor cantidad de valores nulos.

## Gráfico de Nulos en Listings.csv

## Observacion de patron de nulos



A continuacion se muestran los nulos encontrados en el dataset Listings.csv usando una matriz de calor que nos muestra la posicion de estos nulos en el dataset.

## ◆ Identificando nulos

Column	Missing Pattern	NULL %	Action
Description	MCAR	3.8	Flag("Sin descripcion")
Neighborhood_overview	MAR O MNAR	50.7	DELETE
host_location	MAR O MNAR	23.2	DELETE
host_about	MAR O MNAR	37.1	DELETE
host_response_time	MAR	16.1	DELETE
host_response_rate	MAR	16.1	DELETE
host_acceptance_rate	MAR	14.2	DELETE
host_is_superhost	MCAR	2.9	DELETE
host_neighbourhood	MAR O MNAR	51.1	DELETE
neighbourhood	MAR O MNAR	50.7	DELETE
bathrooms	MAR O MNAR	21.3	Parse bathrooms_text
bedrooms	MAR O MNAR	10.2	Regresion bayesiana
beds	MAR O MNAR	21.7	Regresion bayesiana
price	MAR O MNAR	21.4	Regresion bayesiana
has_availability	MCAR	5.6	Imputacion con false
estimated_revenue_1365d	MAR O MNAR	21.4	MICE
ffirst_review	MAR O MNAR	25.3	DELETE
last_review	MAR O MNAR	25.3	DELETE
review_scores_accuracy	MAR	25.3	DELETE
review_scores_cleanliness	MAR	25.3	DELETE
review_scores_checkin	MAR	25.3	DELETE
review_scores_communication	MAR	25.3	DELETE
review_scores_location	MAR	25.3	DELETE
review_scores_value	MAR	25.3	Flag o tratar nulos
license	MAR O MNAR	32.0	DELETE
reviews_per month	MAR	25.3	Flag("Sin descripcion")

Una vez tenemos identificados los nulos, es importante observar el patron de estos nulos. Debido a que teniendo tanto la cantidad de nulos como el patron de estos, podemos tomar decisiones informadas sobre el tratamiento de los datos faltantes.

A continuacion se muestra un archivo de Excel donde me he guiado para identificar los nulos y su patron. Las resoluciones y decisiones que se dicen tomar en el excel son intuitivas y algunas variaran por motivos de contexto que me encontrare mas adelante en el dataset.



# Limpieza de Datos

En esta sección explicamos cómo llevamos a cabo el proceso de limpieza de datos antes del análisis.

Organizamos la limpieza en pasos separados mediante una estructura de carpetas, donde cada subcarpeta representaba una fase del preprocesamiento: eliminación de duplicados, renombrado de columnas, imputación de valores nulos, codificación, etc.



## Estructura de carpetas y pasos previos

- **Eliminación de duplicados:** Se eliminaron registros completamente repetidos utilizando `drop_duplicates()`.
- **Estandarización de nombres de columnas:**
  - Convertimos todos los nombres a minúsculas.
  - Reemplazamos espacios y caracteres especiales por `_`.
  - Eliminamos caracteres no alfabéticos innecesarios.
  - Ejemplo:

```
df.columns = df.columns.str.lower().str.strip().str.replace(' ', '_').str.replace(r'\W+', '', regex=True)
```

- **Eliminación de columnas innecesarias:**

Se eliminaron columnas irrelevantes para el análisis, como identificadores únicos que no aportaban valor, timestamps duplicados, entre otros.

## Tratamiento de valores nulos

Clasificamos los valores nulos en tres categorías según el patrón de *missingness*:

- **MCAR** (Missing Completely At Random): Imputación simple (media, mediana o moda).
- **MAR** (Missing At Random): Imputación con modelos (regresión, MICE, etc.).
- **MNAR** (Missing Not At Random): En general, no imputamos. Creamos flags o eliminamos.

### bathrooms

- Si `bathrooms` está vacío pero existe `bathrooms_text`, se extrae el valor de ahí.
- Si no hay ninguna referencia, se imputa usando la media de baños para alojamientos similares.
- Se añadió una columna `flag` indicando si el valor fue imputado.

### bedrooms, beds

- Variables numéricas continuas con pocos nulos.
- Se usó **regresión bayesiana (BayesianRidge)** para predecir valores usando variables correlacionadas como `accommodates`, `property_type`, etc.

### price

- Se convirtió el precio a formato numérico (`float`).
- Si no había `price` ni `estimated_revenue`, o la disponibilidad era baja, se marcó como cero.
- Resto: imputación mediante regresión bayesiana.
- Se añadió columna `flag_price_imputed` para trazabilidad.

## availability\_30, availability\_365

- Nulos suelen indicar alojamientos inactivos.
- Se imputó con **0** cuando había baja disponibilidad e ingresos.
- En otros casos se mantuvo nulo o se imputó con la mediana (patrón MAR).

## estimated\_revenue\_l365d

- Si price = 0 y availability\_365 = 0 → revenue = 0.
- Si había price y availability pero faltaba revenue, se imputó con **MICE** (Multivariate Imputation by Chained Equations).

## reviews\_per\_month, number\_of\_reviews

- Si el alojamiento nunca recibió una review → imputación con **0**.
- Se añadió columna binaria has\_reviews.

## review\_scores\_\*

- No se imputaron valores nulos en puntuaciones, ya que sin reviews no es posible estimarlas.
- Se mantuvieron nulos.

## Codificación y Flags

- Variables categóricas binarias → **Label Encoding**.
- Variables nominales → Se evitó One-Hot Encoding para no aumentar la dimensionalidad.
- Para cada imputación se añadió una columna flag\_ para identificar valores imputados.

# Detección y tratamiento de Outliers

El tratamiento se hizo por variable:

- En **beds, bedrooms, bathrooms, price, estimated\_revenue**:
  - Se aplicó *capping* dinámico (percentiles P1–P99) para suavizar extremos sin eliminar información.
  - Ejemplo:

```
if use_percentile:  
    lower, upper = df[col].quantile([0.01, 0.99])  
else:  
    Q1 = df[col].quantile(0.25)  
    Q3 = df[col].quantile(0.75)  
    IQR = Q3 - Q1  
    lower = Q1 - multiplier * IQR  
    upper = Q3 + multiplier * IQR
```

- En **number\_of\_reviews, accommodates**:  
Se mantuvieron los outliers ya que reflejan comportamientos reales extremos (ej. hostales con muchas camas).