

DOCUMENTACION del Proyecto:

Limpieza y Transformación de Datos (ETL) con el Dataset de Airbnb

Lucano Riquelme









12/09/25

luciano.luiz.riquelme@gmail.com

1. Objetivo del Proyecto

El objetivo de este proyecto es demostrar un proceso completo de **Extracción, Transformación y Carga (ETL)**. Se tomó un dataset crudo de listados de Airbnb, que presentaba problemas comunes de calidad de datos, y se transformó en un conjunto de datos limpio, estructurado y listo para el análisis.

Pasos Clave del Proceso:

- Cargar datos 
 - Realizar un diagnóstico general del dataframe 
 - Verificar y corregir tipos de datos 
 - Limpiar columnas numéricas y categóricas 
 - Convertir columnas a formato de fecha 
 - Manejar valores nulos (Imputación y eliminación) 
 - Crear nuevas características 
 - Eliminar columnas irrelevantes 
-

2. Extracción: Carga y Diagnóstico Inicial

El primer paso consistió en cargar el dataset listings.csv, que se encontraba comprimido. Se utilizó la librería Pandas en Python para esta tarea.

Código de Carga:

```
import pandas as pd

# Cargar el dataset original de Airbnb

df = pd.read_csv(
    "listings.csv",
    compression='gzip',
    delimiter=";",
    encoding="utf-8",
```

```
thousands=".")
)
```

Una vez cargado, se realizó un diagnóstico inicial para identificar los problemas de calidad.

Resumen Inicial del DataFrame (df.info()):

--- Información ORIGINAL del DataFrame

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

RangeIndex: 18927 entries, 0 to 18926

Data columns (total 79 columns):

...

dtypes: float64(19), int64(24), object(36)

memory usage: 11.4+ MB

Principales Problemas Detectados:

1. **Valores Faltantes:** Columnas críticas como price, reviews_per_month y bedrooms tenían miles de datos faltantes. Otras como neighborhood_overview estaban vacías en más del 50%.
2. **Tipos de Datos Incorrectos:** Columnas que debían ser numéricas como price (ej. "\$232.00") o host_acceptance_rate (ej. "91%") estaban almacenadas como texto (object), impidiendo cualquier cálculo.
3. **Columnas Irrelevantes:** El dataset contenía 79 columnas, muchas de las cuales (URLs, IDs, texto libre) no eran útiles para un análisis cuantitativo y añadían "ruido".



3. Transformación: El Proceso de Limpieza

Para solucionar los problemas identificados, se aplicaron una serie de técnicas de limpieza y transformación de datos.

Pasos Realizados:

- **Limpieza de Columnas Numéricas:** Se eliminaron símbolos monetarios (\$), porcentajes (%) y comas de las columnas price, host_response_rate y host_acceptance_rate, y se convirtieron a tipo numérico (float).
- **Conversión de Fechas:** Las columnas last_scraped y host_since se convirtieron a formato datetime para permitir análisis temporales.
- **Manejo de Nulos:**
 - Se eliminaron columnas con más del 50% de valores faltantes.
 - Se rellenaron (imputaron) los valores nulos en columnas numéricas clave utilizando la **mediana**.

- Se rellenaron los nulos en columnas categóricas con el valor más frecuente (la **moda**).
- Se eliminaron las 3 únicas filas que tenían un valor nulo irreparable en `host_since`.
- **Ingeniería de Características:** Se creó una nueva columna `num_amenities` contando el número de servicios ofrecidos en cada listado a partir de la columna de texto `amenities`.
- **Eliminación de Columnas:** Se eliminaron 15 columnas redundantes o innecesarias para enfocar el dataset en la información más relevante.

✅ 4. Carga: El Resultado Final

Tras el proceso de limpieza, el dataset quedó completamente transformado y listo para ser cargado en una base de datos o utilizado para análisis y visualización.

Resumen Final del DataFrame Limpio (`df_limpio.info()`):

--- Información final del DataFrame limpio ---

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

Index: 18924 entries, 0 to 18926

Data columns (total 64 columns):

...

dtypes: bool(2), datetime64[ns](2), float64(21), int64(24), object(15)

memory usage: 9.1+ MB

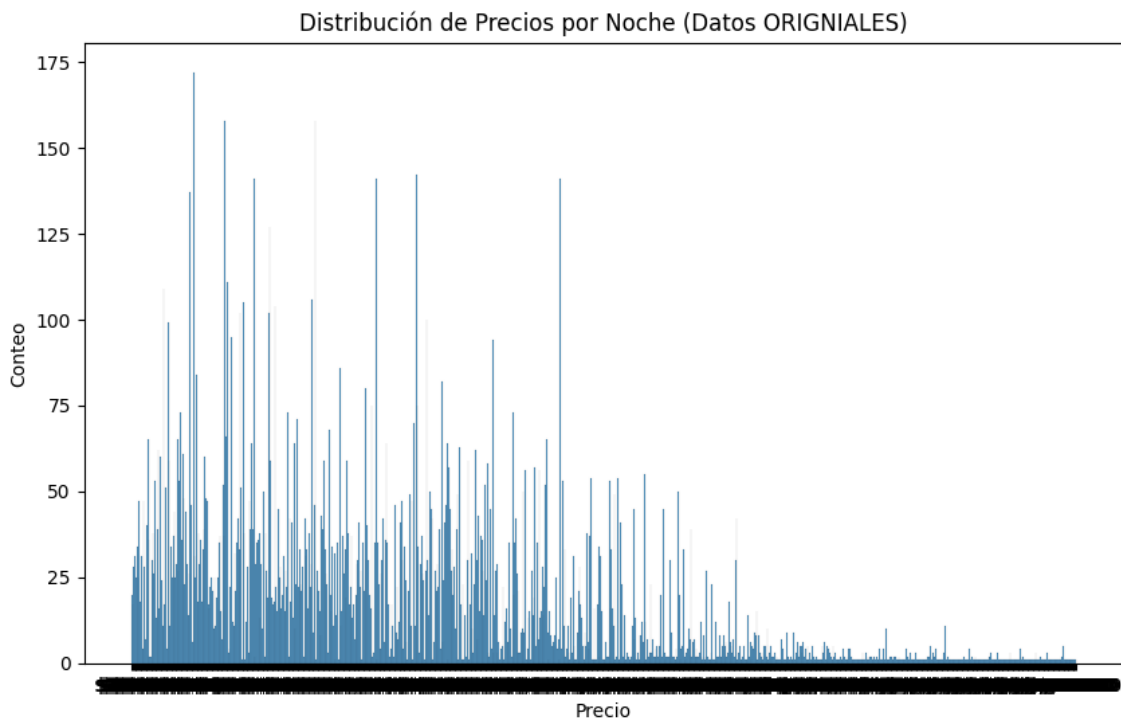
Comparación del Antes y Después:

Característica	Dataset Original (Antes)	Dataset Limpio (Después)
Número de Filas	18,927	18,924
Número de Columnas	79	64
Valores Nulos Totales	Decenas de miles	CERO
Columnas de Texto (object)	36	15

📊 5. Demostración Visual del Impacto

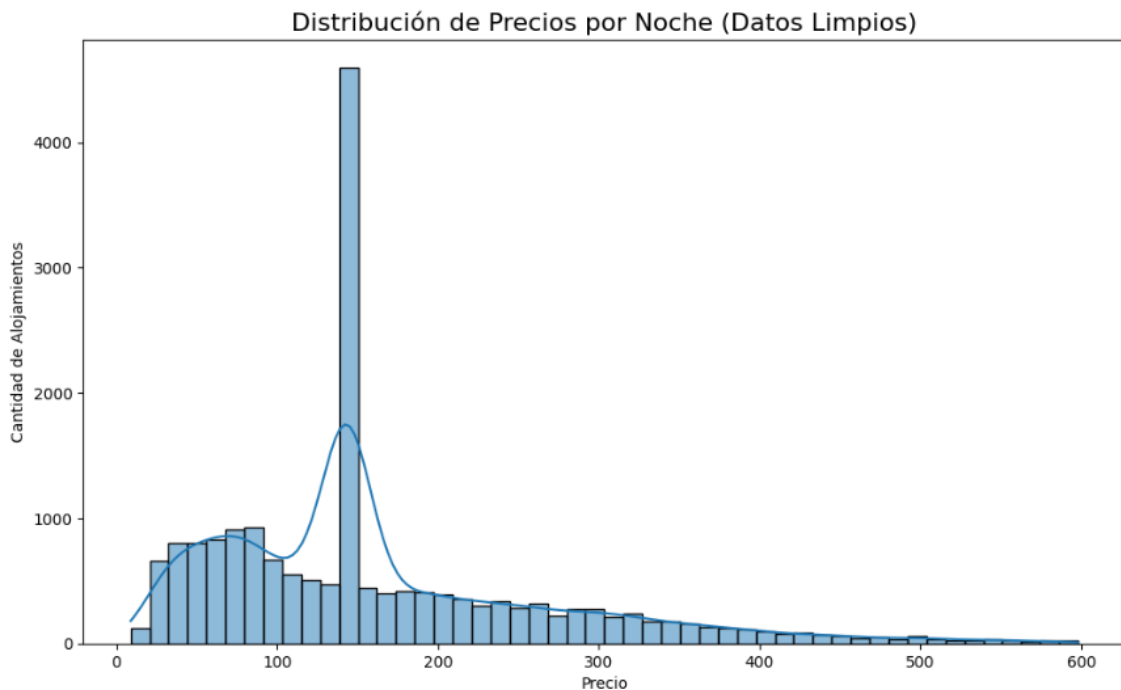
La mejor forma de demostrar la utilidad de la limpieza de datos es a través de la visualización. Con los datos originales, era imposible generar gráficos numéricos sobre el precio.

Gráfico "Antes": Intento de Graficar con Datos Sucios



Observación: El gráfico demuestra que la columna de precios original, al ser texto, no permitía ningún tipo de análisis de distribución.

Gráfico "Después": Distribución de Precios con Datos Limpios



Observación: Una vez limpia, la misma columna nos permite visualizar claramente que la mayoría de los alojamientos se concentran en el rango de precios de \$50 a \$200 por noche. **Este análisis era imposible antes del proceso de ETL.**

6. Conclusión

Este proyecto demuestra la capacidad de tomar un conjunto de datos crudo y desordenado y, a través de un proceso sistemático de ETL, convertirlo en un activo de datos valioso, íntegro y fiable. El dataset final está optimizado y completamente preparado para análisis exploratorio, la creación de dashboards o el entrenamiento de modelos de machine learning.