



TALLER 2

Proceso ETL con datasets de Airbnb Ciudad de Mexico



26 DE OCTUBRE DE 2025

YENIFER GONZALEZ QUIRAMA

EMILIANO VELEZ SUAREZ

Introducción

El presente informe describe el desarrollo de un proceso **ETL (Extracción, Transformación y Carga)** aplicado sobre los datasets de **Airbnb Ciudad de México**, con el propósito de construir una base de datos limpia y estructurada que sirva como base para análisis y toma de decisiones.

El proyecto se desarrolló en **Python**, empleando librerías como **pandas**, **sqlite3**, **logging** y **openpyxl**, siguiendo las tres fases fundamentales del proceso ETL: **Extracción** de datos desde MongoDB, **Transformación** mediante limpieza y enriquecimiento, y **Carga** en SQLite y Excel.

Descripción del dataset

El dataset proviene de **Airbnb Ciudad de México** y contiene tres colecciones principales:

- **listings:** Información de los alojamientos (anfitrión, tipo de habitación, precio, ubicación).
- **calendar:** Registros diarios de disponibilidad y precio.
- **reviews:** Opiniones y valoraciones de los huéspedes.

Los datos fueron exportados desde MongoDB a formato **CSV** para su procesamiento.

Análisis exploratorio

El análisis exploratorio se realizó utilizando **Python y Jupyter Notebook**, con el objetivo de identificar patrones generales, posibles inconsistencias y variables relevantes para el proceso ETL.

Durante esta fase se inspeccionaron los tres datasets principales (listings, calendar, reviews), revisando su estructura, cantidad de registros y tipos de datos. Se verificó que **no existían valores duplicados significativos ni nulos relevantes**, por lo que no fue necesario aplicar procesos extensivos de limpieza en ese aspecto.

En cambio, se enfocó el trabajo en:

- **Conversión de tipos de datos:**
Se estandarizaron las fechas al formato YYYY-MM-DD y los precios a valores numéricos, eliminando símbolos como \$ y ,.

- **Normalización de texto:**
Se homogenizaron los nombres de colonias y tipos de propiedad (room_type) para evitar diferencias por mayúsculas o acentos.
- **Distribución de tipos de alojamiento:**
Se encontró que el tipo de propiedad más frecuente es “**Entire home/apt**” (más del 70% de los registros), seguido por “**Private room**”
- **Detección de Outliers**
Se identifican valores atípicos en el precio (> \$1,000 por noche).

Los principales hallazgos fueron los siguientes:

- No se encontraron registros duplicados en ninguno de los tres datasets.
- No existían valores nulos relevantes en las columnas principales (id, listing_id, price, date).
- Se detectaron campos con estructura anidada, como host_verifications y amenities, los cuales contenían listas dentro de una sola celda.
- Algunos campos de texto requerían normalización en los nombres de columnas y valores, especialmente los relacionados con el tipo de habitación y el vecindario.
- Se verificó que el campo price incluía el símbolo \$ y comas, lo cual fue corregido durante la etapa de transformación.

Descripción de las transformaciones realizadas

Tipo de transformación	Descripción
Limpieza de datos	Eliminación de valores nulos irrelevantes y verificación de integridad de columnas. No se encontraron duplicados.
Normalización de precios	Se eliminaron símbolos \$ y , para convertir la columna price a valores numéricos.
Conversión de fechas	Se transformaron al formato estándar ISO YYYY-MM-DD.

Tipo de transformación	Descripción
Derivación de variables	A partir de date se extrajeron las columnas de mes, año y trimestre.
Desanidado de campos	se expandieron los campos host_verifications y amenities en columnas binarias, representando la presencia o ausencia de cada elemento.
Creacion de variable derivada	Se generó el campo reviews_per_month con base en la relación entre los datasets reviews y calendar, permitiendo calcular la cantidad promedio de reseñas mensuales por alojamiento.

Todas las operaciones fueron registradas en logs para trazabilidad del proceso.

Ejemplo del log generado

2025-10-18 08:22:31 - INFO - Inicio del proceso ETL.

2025-10-18 08:22:32 - INFO - Registros iniciales: 54764

2025-10-18 08:22:35 - INFO - Valores nulos eliminados correctamente.

2025-10-18 08:22:37 - INFO - Precios normalizados y convertidos a float.

2025-10-18 08:22:40 - INFO - Fechas convertidas al formato ISO.

2025-10-18 08:22:45 - INFO - Variables derivadas agregadas: mes, año, trimestre.

2025-10-18 08:22:50 - INFO - Registros cargados correctamente en airbnb.db (tabla calendar_limpio).

2025-10-18 08:22:52 - INFO - Archivo Excel generado: airbnb_limpio.xlsx

2025-10-18 08:22:52 - INFO - Proceso ETL finalizado exitosamente.

Conclusiones sobre la calidad y utilidad de los datos

- Los datos analizados presentan alta consistencia y completitud, sin duplicados relevantes ni valores faltantes críticos.

- Las transformaciones realizadas permitieron mejorar la estructura de los datos para su análisis posterior en herramientas como Power BI o Excel
- La inclusión de variables derivadas (mes, año, rango de precios) facilitó la generación de indicadores y gráficos descriptivos.
- El dataset de Airbnb para Ciudad de México se considera fiable y útil para análisis exploratorios y predictivos, permitiendo obtener una visión clara del comportamiento del mercado de hospedaje.

Referencias

- Inside Airbnb – Dataset Ciudad de México: <https://insideairbnb.com/get-the-data/>
- Documentación oficial de pandas: <https://pandas.pydata.org>
- Python Software Foundation – módulo sqlite3
- Guías académicas del curso