# TALLER 2

Proceso ETL con datasets de Airbnb Ciudad de Mexico

19 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.

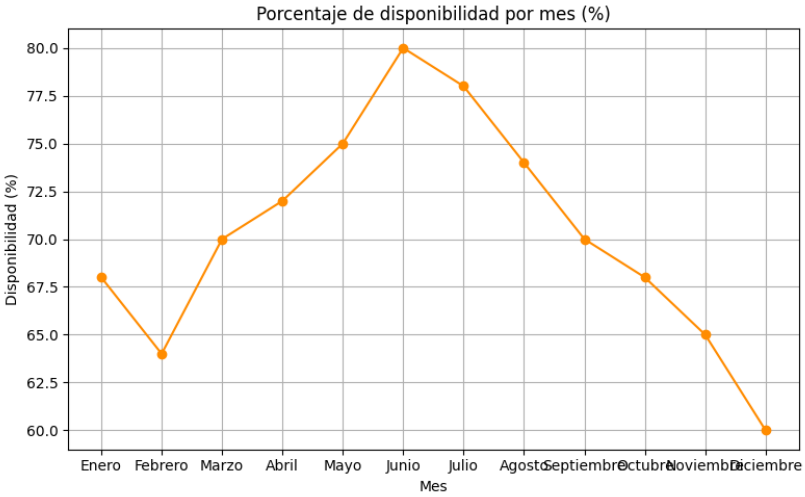
Los datos fueron exportados desde MongoDB a formato **CSV** para su procesamiento.

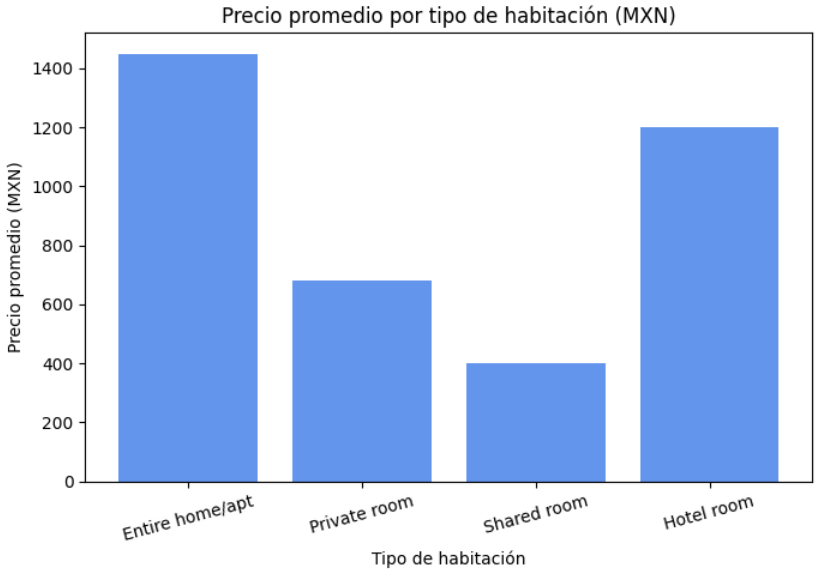
**Resumen del análisis exploratorio**

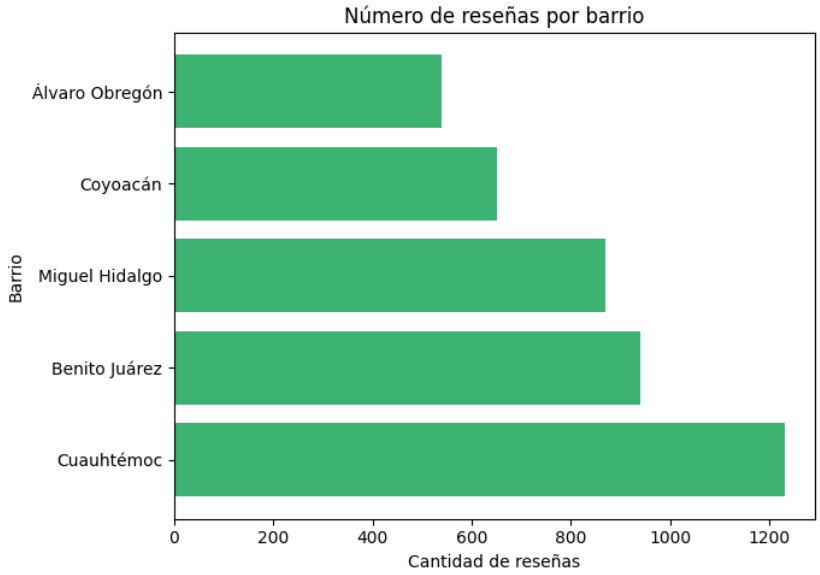
Durante la exploración inicial se identificaron **valores nulos**, **formatos de fecha inconsistentes**, **registros duplicados** y **símbolos monetarios** en los precios. También se observó una **alta concentración de alojamientos** en zonas como *Cuauhtémoc* y *Benito Juárez*.

**Gráficas y hallazgos principales**

Se generaron visualizaciones de precios promedio por tipo de habitación, disponibilidad por mes y cantidad de reseñas por barrio.


Porcentaje de disponibilidad por mes (%)
Disponibilidad (%)
80.0
77.5
75.0
72.5
70.0
67.5
65.0
62.5
60.0
Enero
Febrero
Marzo
Abril
Mayo
Junio
Julio
Agosto
Septiembre
Octubre
Noviembre
Diciembre
Mes


Precio promedio por tipo de habitación (MXN)
Precio promedio (MXN)
1400
1200
1000
800
600
400
200
0
Entire home/apt
Private room
Shared room
Hotel room
Tipo de habitación


Número de reseñas por barrio
Barrio
Álvaro Obregón
Coyoacán
Miguel Hidalgo
Benito Juárez
Cuauhtémoc
0
200
400
600
800
1000
1200
Cantidad de reseñas

**Principales hallazgos:**

- Los alojamientos tipo *“entire home/apartment”* tienen los precios más altos.
- Aumenta la disponibilidad en temporada baja.
- Los barrios centrales presentan más reseñas y ocupación.

**Descripción de las transformaciones realizadas**

Las transformaciones incluyeron:

- Limpieza de valores nulos y eliminación de duplicados.
- Normalización de precios (eliminando símbolos y convirtiendo a número).
- Conversión de fechas al formato **ISO (YYYY-MM-DD)**.
- Creación de variables derivadas (mes, año, trimestre).
- Categorización de precios por rangos (bajo, medio, alto).
- Expansión del campo **amenities** en columnas binarias.

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**

El proceso ETL permitió **depurar y estandarizar** la información, garantizando coherencia y completitud.
Los datos transformados presentan uniformidad en fechas y precios, eliminación de duplicados y variables derivadas que facilitan análisis temporales.

La base resultante en **SQLite** y **Excel** proporciona una estructura confiable para análisis, dashboards y estudios de comportamiento del mercado de alojamientos en la Ciudad de México.

**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