Informe Final

Análisis Exploratorio y Transformaciones — Airbnb Ciudad de México

Integrantes:

Santiago Varela Jiménez
Santiago Álzate Munera
Franyelica García Fernández

2025-10-24

2. Introducción

Objetivo: Presentar el análisis exploratorio realizado sobre las colecciones extraídas de la plataforma Airbnb (Ciudad de México), detallar las transformaciones aplicadas para limpieza y normalización, y evaluar la calidad de los datos para su posterior carga en un Data Warehouse.

Alcance: Se analizan las colecciones MX_listings, MX_reviews y MX_calendar; se documentan hallazgos relevantes, transformaciones aplicadas y ejemplos de logs.

3. Descripción del dataset

Colecciones analizadas:
- MX_listings (listings): información de alojamientos — 26,401 registros y 77 columnas (según EDA)
- MX_reviews (reviews): reseñas de usuarios — 1,388,226 registros y 7 columnas
- MX_calendar (calendar): disponibilidad por fecha — 9,636,365 registros y 6 columnas

Columnas clave (ejemplos):
- listings: id, name, host_id, host_name, neighbourhood, latitude, longitude, price, minimum_nights, availability_365, amenities, host_verifications, review_scores_*.
- reviews: listing_id, id, date, reviewer_id, reviewer_name, comments.
- calendar: listing_id, date, available, price.

Fuente y extracción: Datos extraídos desde MongoDB (MX_DB) utilizando una conexión local y scripts en Python (colecciones exportadas con la clase Extraccion).

Nota: Los datos reales del dataset de Airbnb (Ciudad de México) fueron recortados; se está trabajando únicamente con el último trimestre. De esta manera se logra un análisis más preciso y una carga de datos más eficiente.

4. Resumen del análisis exploratorio (EDA)

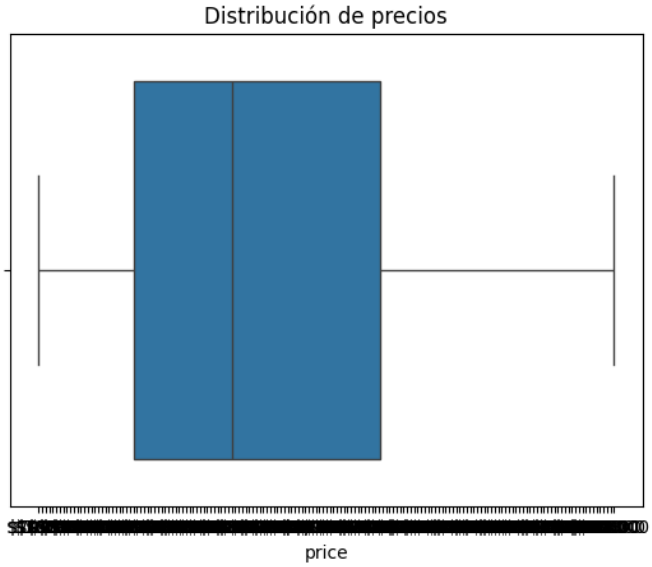
Estadísticas y observaciones principales:

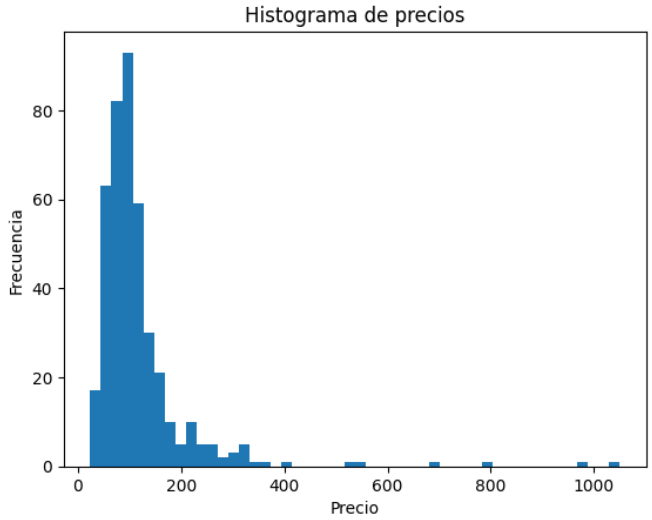
Precios: la columna price contenía símbolos ($) y estaba almacenada como object; tras limpieza se convirtió a numérico. Distribución sesgada hacia precios bajos con algunos outliers altos.
Minimum_nights: presencia de valores extremos (algunos > 1000) que requieren revisión o truncamiento según reglas de negocio.
Disponibilidad: análisis por mes muestra variación estacional; meses como junio-julio presentan menor disponibilidad en los datos analizados.
Valores nulos: listings contiene nulos en campos descriptivos (host_about, neighbourhood_overview, review_scores_*). reviews y calendar no presentan nulos significativos en columnas clave.
Duplicados: no se detectaron duplicados significativos después de convertir columnas no hasheables a string para el conteo de duplicados.
Campos anidados: amenities y host_verifications son listas; amenities puede expandirse en variables binarias o contar elementos para análisis.

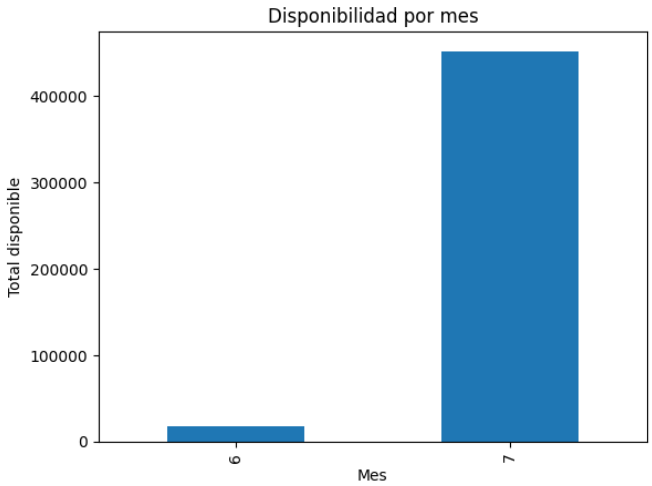
5. Gráficas y hallazgos principales

Gráficas incluidas (referenciar imágenes en el PDF final):
- Boxplot de price: muestra distribución y outliers.
- Histograma de price: concentración de precios bajos.
- Gráfico de barras: disponibilidad total por mes (a partir de calendar).
- Conteo de amenities más frecuentes (si se desanida y se analiza).

Interpretación breve:
- La mayoría de los alojamientos tienen precios bajos; los outliers pueden indicar propiedades de lujo o errores de registro.
- Patrón estacional detectado en disponibilidad que puede relacionarse con demanda turística.


Distribución de precios
price

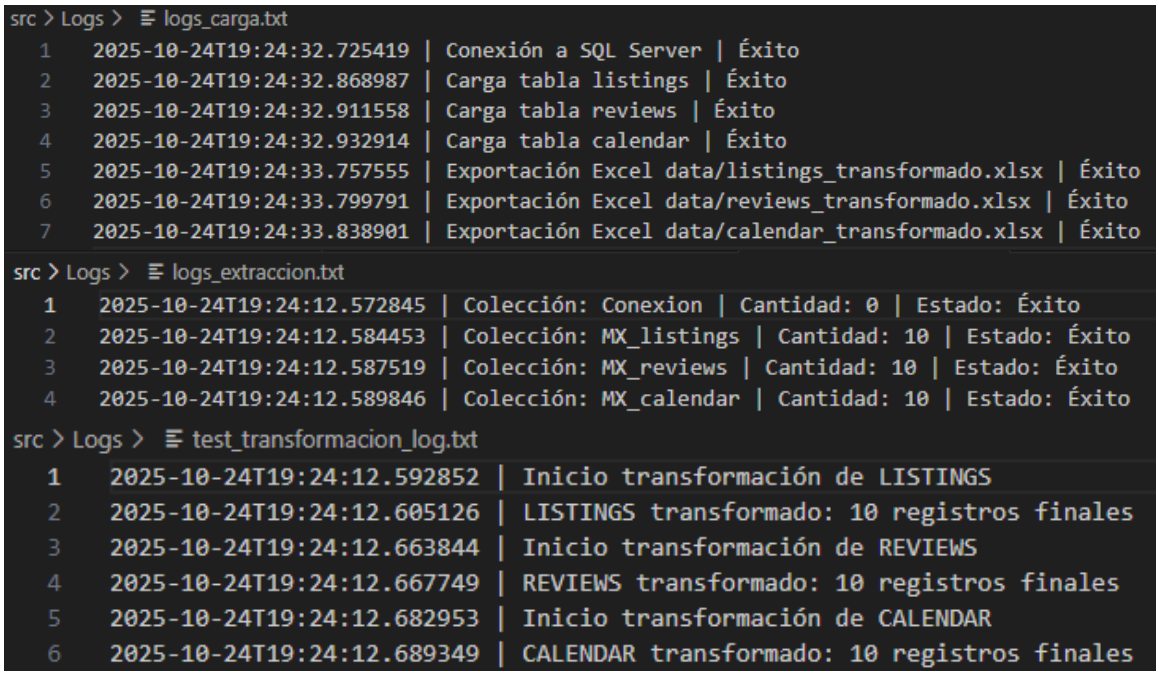

Histograma de precios
Frecuencia
0
20
40
60
80
0
200
400
600
800
1000
Precio


Disponibilidad por mes
Total disponible
0
100000
200000
300000
400000
6
7
Mes

6. Descripción de las transformaciones realizadas

Pasos aplicados:
- Conversión de tipos:
  * calendar.date → datetime
  * price (en listings y calendar) → limpiar símbolos y convertir a float
- Limpieza de price (ejemplo):
  Antes: "$1,200.00" → Después: 1200.0
  Código usado (ejemplo): df['price'] = df['price'].astype(str).replace(r'[\$,]', '', regex=True).replace('', '0').astype(float)
- Manejo de campos anidados:
  * amenities: almacenado originalmente como lista; se creó columna amenities_count con el número de elementos y se dejó la lista como JSON/string para análisis posterior.
- Valores nulos y extremos:
  * Imputación: se dejaron nulos en columnas descriptivas y se documentaron (no imputadas por defecto).
  * Outliers: se identificaron y se marcaron para revisión; no se eliminaron automáticamente.
- Duplicados: se buscó duplicidad por filas completas tras convertir columnas no hasheables a texto; no se eliminaron registros masivamente porque no se detectaron duplicados confiables.

7. Ejemplo del log generado



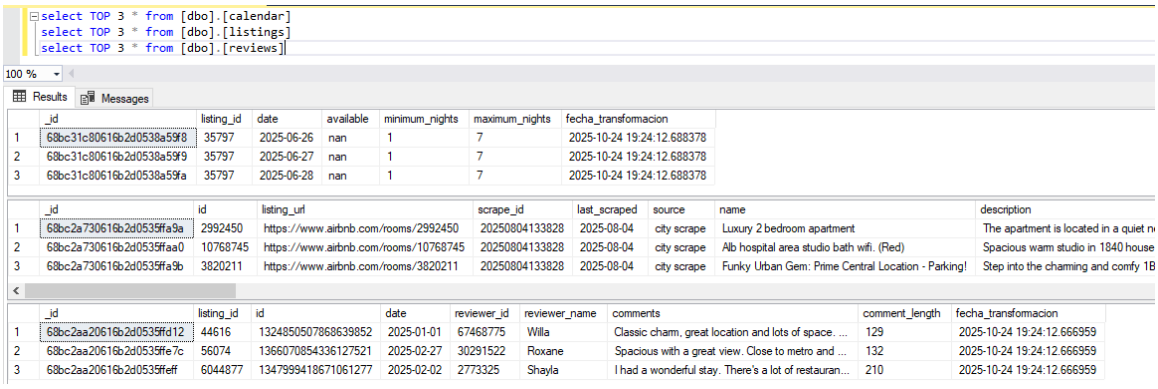
8. Conclusiones sobre la calidad y utilidad de los datos

Fortalezas:
- Volumen y granularidad adecuados para análisis descriptivo y modelado (especialmente con reviews y calendar).
- Fechas en formato susceptible de análisis temporal después de conversión.

Limitaciones:
- Campos descriptivos con valores nulos que limitan análisis semánticos sin imputación o enriquecimiento.
- Presencia de outliers en price y minimum_nights que requieren reglas de negocio para limpieza.
- Columnas anidadas (listas) que requieren transformación para algunos modelos.

Recomendaciones:
- Definir reglas de negocio para manejar outliers y valores extremos.
- Enriquecer datos geográficos (reverse geocoding) si se requiere análisis por zona.
- Expandir amenities a variables binarias si se va a usar para modelado.

9. Conexion DB SQL Server



```
select TOP 3 * from [dbo].[calendar]
select TOP 3 * from [dbo].[listings]
select TOP 3 * from [dbo].[reviews]
```

100 %

Results | Messages

	_id	listing_id	date	available	minimum_nights	maximum_nights	fecha_transformacion
1	68bc31c80616b2d0538a59f8	35797	2025-06-26	nan	1	7	2025-10-24 19:24:12.688378
2	68bc31c80616b2d0538a59f9	35797	2025-06-27	nan	1	7	2025-10-24 19:24:12.688378
3	68bc31c80616b2d0538a59fa	35797	2025-06-28	nan	1	7	2025-10-24 19:24:12.688378

	_id	id	listing_url	scrape_id	last_scraped	source	name	description
1	68bc2a730616b2d0535ffa9a	2992450	https://www.airbnb.com/rooms/2992450	20250804133828	2025-08-04	city scrape	Luxury 2 bedroom apartment	The apartment is located in a quiet n
2	68bc2a730616b2d0535ffaa0	10768745	https://www.airbnb.com/rooms/10768745	20250804133828	2025-08-04	city scrape	Alb hospital area studio bath wifi. (Red)	Spacious warm studio in 1840 house
3	68bc2a730616b2d0535ffa9b	3820211	https://www.airbnb.com/rooms/3820211	20250804133828	2025-08-04	city scrape	Funky Urban Gem: Prime Central Location - Parking!	Step into the charming and comfy 1B

	_id	listing_id	id	date	reviewer_id	reviewer_name	comments	comment_length	fecha_transformacion
1	68bc2aa20616b2d0535ffd12	44616	1324850507868639852	2025-01-01	67468775	Willa	Classic charm, great location and lots of space. ...	129	2025-10-24 19:24:12.666959
2	68bc2aa20616b2d0535ffe7c	56074	1366070854336127521	2025-02-27	30291522	Roxane	Spacious with a great view. Close to metro and ...	132	2025-10-24 19:24:12.666959
3	68bc2aa20616b2d0535ffeff	6044877	1347999418671061277	2025-02-02	2773325	Shayla	I had a wonderful stay. There's a lot of restauran...	210	2025-10-24 19:24:12.666959

10. Referencias

- Fuente de datos: Colecciones extraídas desde MongoDB (instancia local).
- Herramientas y bibliotecas: Python, pandas, seaborn, matplotlib, pymongo (u otro cliente MongoDB).