Tecnológico de Monterrey

Unidad de formación:

Plataforma de analítica de negocios para organizaciones

Actividad 5 -Reporte

Profesor:

Alfredo García Suarez

Alumna:

Pilar Vaquero Fernández

Fecha de entrega:

05 de septiembre de 2025

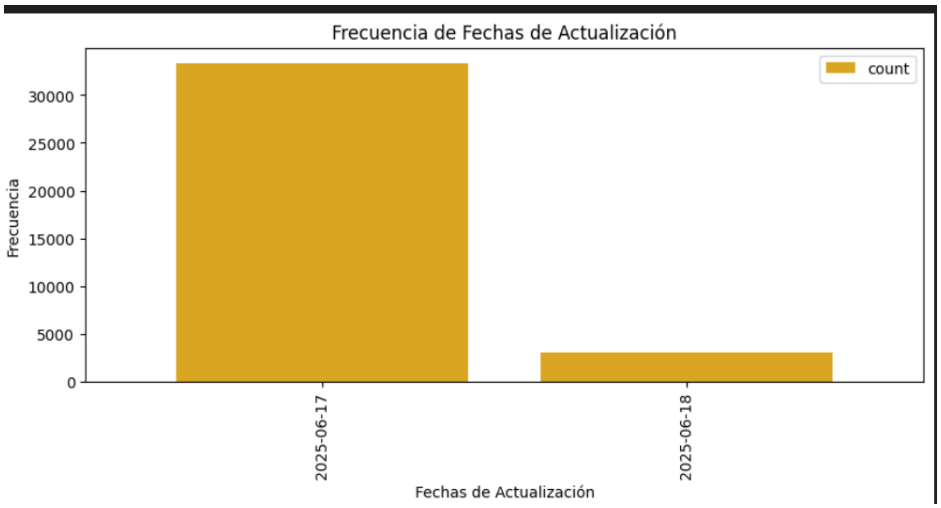
Introducción

Para esta actividad trabajé con el dataset limpio de anuncios en Nueva York. Me concentro en leer proporciones, rangos y lo que implican para el comportamiento de los listados.

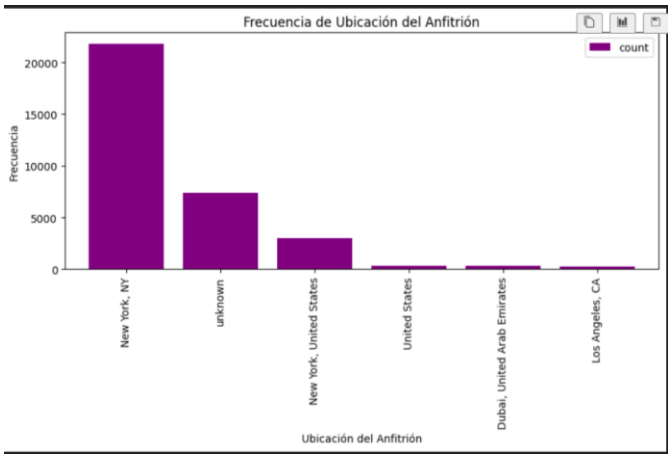
Hice analisis univariados, despues filtros para los valores mas relevantes, asi despues los ajuste y ya cree los graficos para que no tuvieran valores que no fueran relevantes.

1) Frecuencia de Fechas de Actualización (last_scraped) — Gráfico de barras

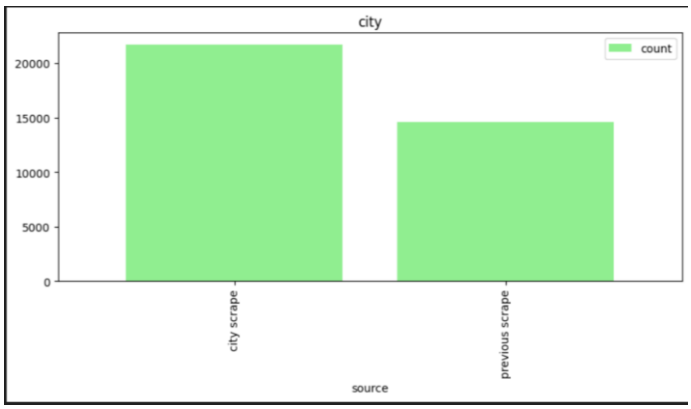
Qué es: conté cuántos anuncios fueron actualizados en cada fecha disponible del campo last_scraped.
Grafico: el gráfico muestra dos barras. La del 2025-06-17 concentra 33,318 anuncios y la del 2025-06-18 tiene 3,004.
Interpretación: la gran mayoría de los registros (≈ 91.7%) se raspó el 17 de junio, y un 8.3% el 18. Es un patrón típico cuando se hace una extracción masiva en un mismo día y se termina al día siguiente. Me confirma que no hay sesgo temporal importante dentro del corte: los datos son esencialmente de una misma ventana.



2)Ubicación del anfitrión (host_location) — Barras
Qué es: conté cuántos anuncios hay por ubicación declarada en host_location (Top categorías).
Gráfico: domina “New York, NY”; en 2.º lugar “unknown”; luego etiquetas genéricas como “New York, United States” y algunas ciudades/países con volúmenes mucho menores.
Interpretación: la oferta es mayoritariamente local (NYC). El peso de “unknown” y variantes sugiere inconsistencia en el dato; conviene normalizar ubicaciones (estandarizar nombres y agrupar “otros”).



3) Origen de los datos (source) — Barras
Qué es: conté cuántos anuncios provienen de cada origen en source (p. ej., *city scrape* vs *previous scrape*).
Gráfico: aparecen dos barras; predomina “city scrape” y “previous scrape” aporta el resto (aprox. 60/40).
Interpretación: el set combina una extracción reciente con un arrastre previo. Útil para control de calidad: revisar posibles duplicados entre scrapes y, si modelamos, considerar source como variable de control por efectos de mezcla temporal.

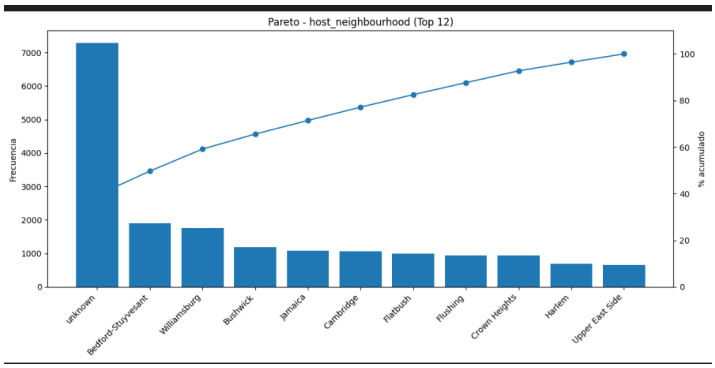


4) Pareto — host_neighbourhood

Qué es: un diagrama de Pareto con las frecuencias por barrio y la línea de porcentaje acumulado.

Gráfico: domina “unknown” (~7.3K), luego Bedford-Stuyvesant (~1.9K) y Williamsburg (~1.75K); el resto cae rápidamente. La línea acumulada cruza ~80% alrededor del 7º barrio.

Interpretación: la concentración es alta: pocos barrios (≈7) explican ≈80% de los hosts. Conviene priorizar esos vecindarios y depurar “unknown” (puede esconder geos clave o errores de captura).

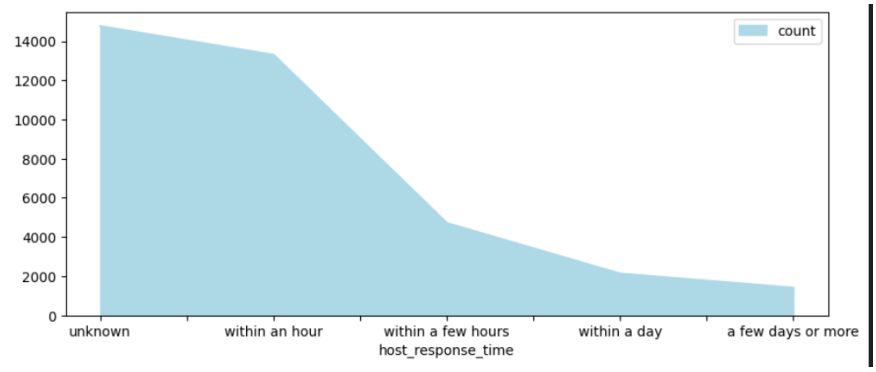


5) host_response_time — Gráfico de área

Qué es: distribución de tiempos de respuesta declarados por los anfitriones.

Gráfico: domina “unknown” (~15 mil); luego within an hour (13,307), within a few hours (4,710), within a day (2,126) y a few days or more (1,407).

Interpretación: alrededor de ~18 mil listados responden ≤ “few hours” (rápidos), unos ~3.5 mil tardan ≥ 1 día, y el gran bloque “unknown” (~40%) sugiere metadatos faltantes; al limpiarlo, la mayoría queda con respuesta ≤ 1 hora.

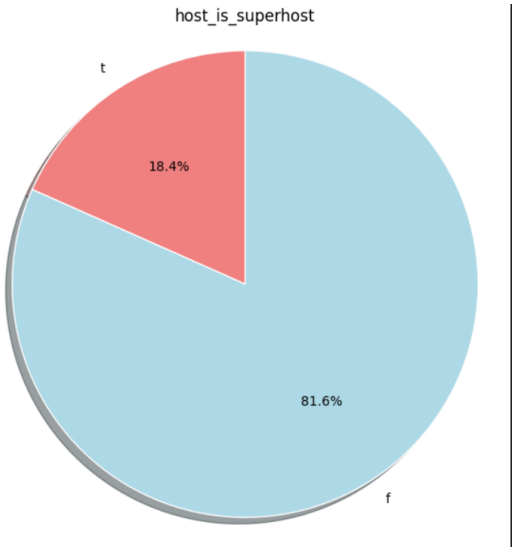


6) host_is_superhost — Gráfico de pastel

Qué es: proporción de anuncios cuyo anfitrión tiene distintivo Superhost.

Gráfico: dos segmentos: f = 81.6% (no superhost) y t = 18.4% (sí superhost).

Interpretación: alrededor de 1 de cada 5 anuncios pertenece a Superhosts; la mayoría (≈4 de cada 5) no tiene ese estatus. Esto sugiere un mercado dominado por anfitriones estándar, con un nicho relevante de Superhosts que podrían mostrar mejor desempeño (p. ej., tasas de respuesta y evaluación), útil para comparativas.

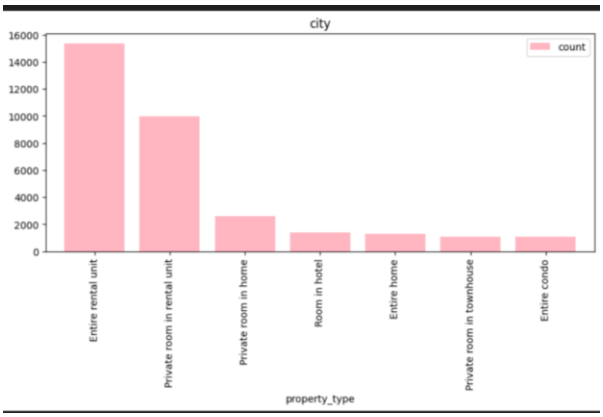


7) Tipos de propiedad (property_type) — Barras

Qué es: conté cuántos anuncios hay por tipo de propiedad.

Gráfico: destacan Entire rental unit (~15.5k) y Private room in rental unit (~10k); luego Private room in home (~2.6–2.8k). El resto (room in hotel, entire home, private room in townhouse, entire condo) está < 1.5k cada uno.

Interpretación: el inventario se concentra en unidades completas y cuartos privados en edificios de renta; los tipos “home/condo/hotel” son minoritarios. Con esas tres primeras categorías se cubre >80% de la oferta, así que las usaría para segmentar precios y ocupación.

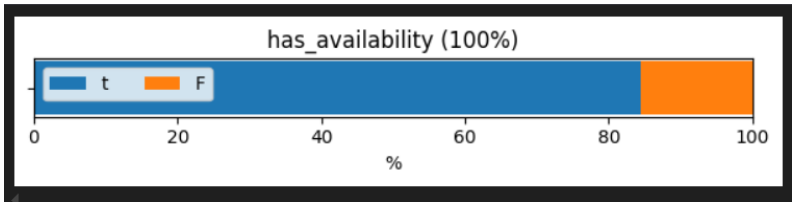


8) has_availability (100%)

Qué es: Muestra la proporción de listings disponibles ('t') vs no disponibles ('F') en calendario.

Gráfico: Barra apilada, aprox. 84.5% 't', 15.5% 'F'.

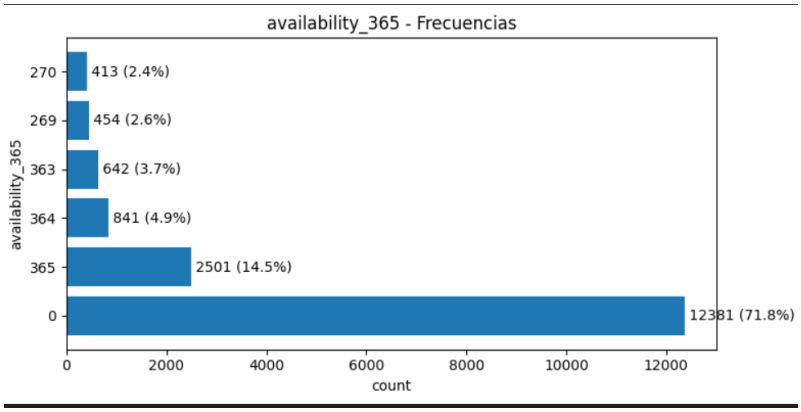
Interpretación: La mayoría de los listings están disponibles, mientras que alrededor del 15% no lo están (por estar vendidos o bloqueados).



9) Qué es: conté cuántos anuncios tienen cada valor de availability_365 (número de días disponibles en el año).

Gráfico: la barra más grande es 0 días con 12,381 anuncios (71.8%). Luego 365 días con 2,501 (14.5%). Las demás (364, 363, 269, 270) son mucho menores: 841 (4.9%), 642 (3.7%), 454 (2.6%) y 413 (2.4%).

Interpretación: la mayoría aparece sin disponibilidad futura (0), lo que puede indicar calendarios cerrados, listados inactivos o ya reservados. Un segundo bloque relevante es el de disponibilidad total (365), que suele ser la oferta más flexible. El resto son casos residuales. Con esto, yo priorizaría:

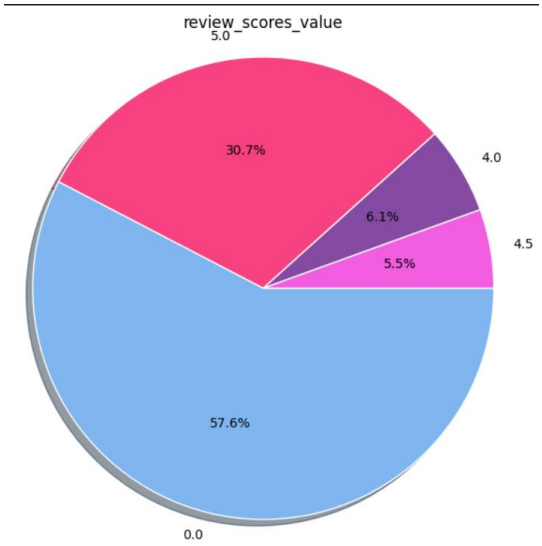


10) review_scores_value — Pastel

Qué es: distribución del puntaje review_scores_value (escala 0–5; 0 suele significar "sin reseñas").

Gráfico: 0.0 = 57.6%, 5.0 = 30.7%, 4.0 = 6.1%, 4.5 = 5.5%.

Interpretación: más de la mitad no tiene valoraciones aún (0.0), así que no son "malas" reseñas sino ausencia de reviews. Entre los que sí tienen, domina la máxima calificación (5.0); los valores 4.0–4.5 son minoría. Yo segmentaría campañas: 1) captar primeras reseñas para el bloque 0.0 y 2) destacar/listar

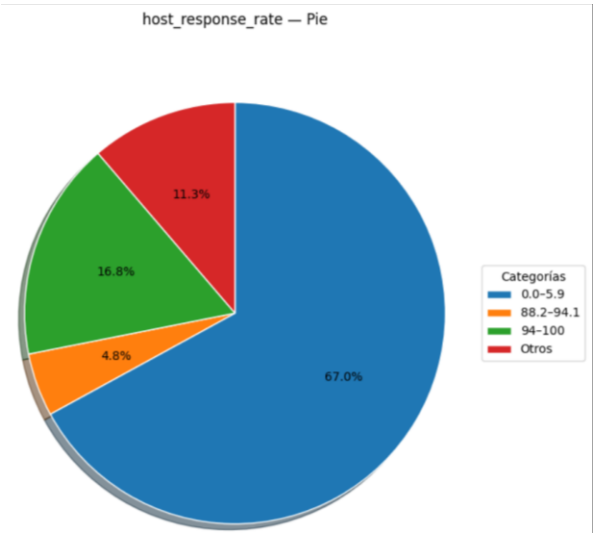


Despues de esto elegí 10 variables clave y confirmé que existieran en el DataFrame; si faltaba alguna, la reemplacé por otra numérica hasta completar diez. Cuando fue necesario, generé variables derivadas (por ejemplo, number_of_reviews_l365d desde *ltm* o *l30d*×12 y reviews_per_month = *ltm*/12) y forcé todo a numérico para evitar errores. Convertí bathrooms_text a un valor decimal (_bathrooms_num) para analizar sin textos. Después apliqué la regla de Sturges para definir el número de intervalos y segmenté cada variable con pd.cut. Con esos cortes construí tablas de frecuencia y porcentaje (count y pct), las guardé en tablas[col] y revisé con head(10); si alguna columna ya existía, mostré un aviso y no la recreé.

1) Qué es: repartí host_response_rate en intervalos y armé un pastel con los tramos más representativos.

Gráfico: cuatro porciones: 0.0–5.9 (≈67.0%), 94–100 (≈16.8%), 88.2–94.1 (≈4.8%) y Otros (≈11.3%).

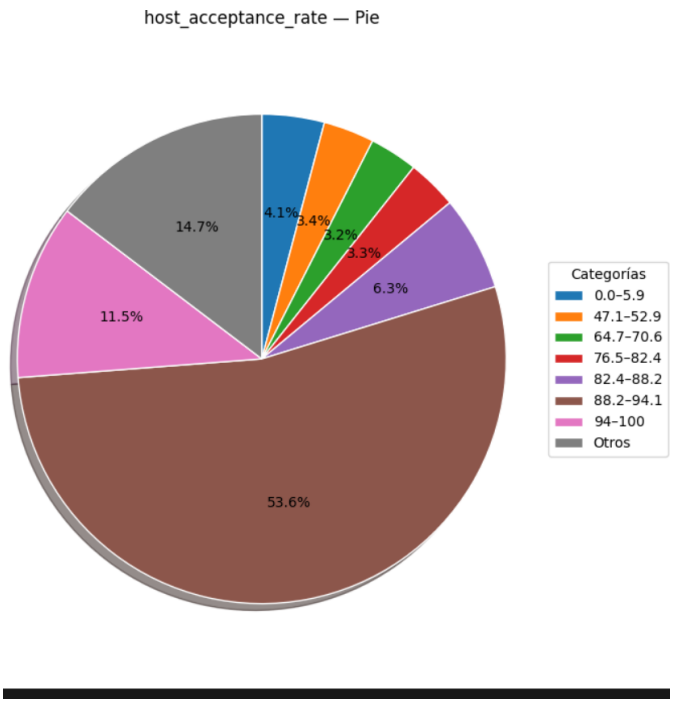
Interpretación: la distribución es muy sesgada: la mayoría cae en el tramo más bajo (0–5.9), pero existe un bloque relevante con tasas altas (94–100). Esto sugiere mezcla de anuncios con nula/bajísima respuesta y anfitriones muy reactivos. Recomiendo verificar si 0% codifica faltantes o “no aplica”; si es así, conviene limpiarlos antes de concluir.



2) Qué es: dividí host_acceptance_rate en rangos y armé un pastel para ver su distribución.

Gráfico: domina el tramo 94–100% (≈53.6%). Le siguen 88.2–94.1% (≈11.5%), 82.4–88.2% (≈6.3%) y varios tramos bajos (≈3–4% cada uno). El grupo “Otros” concentra ≈14.7% (resto de categorías/valores raros).

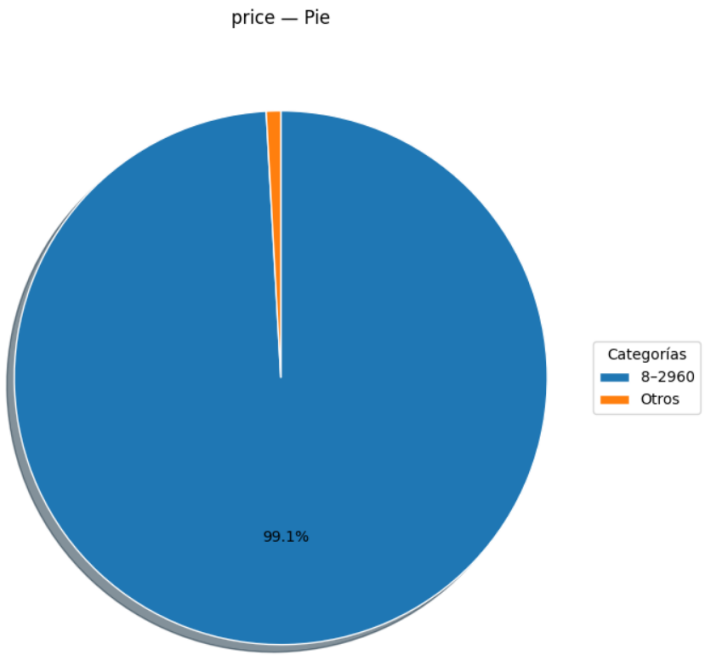
Interpretación: más de la mitad de los anfitriones acepta casi todas las solicitudes; otro ~12% también es alto. El ~15% en “Otros” sugiere valores dispersos o atípicos; conviene revisar si hay faltantes/errores en los rangos bajos antes de sacar conclusiones operativas.



3) Qué es: clasifiqué price en rangos y armé un pastel para ver la distribución.

Gráfico: la categoría 8–2960 concentra ≈99.1% de los anuncios; “Otros” es ≈0.9%.

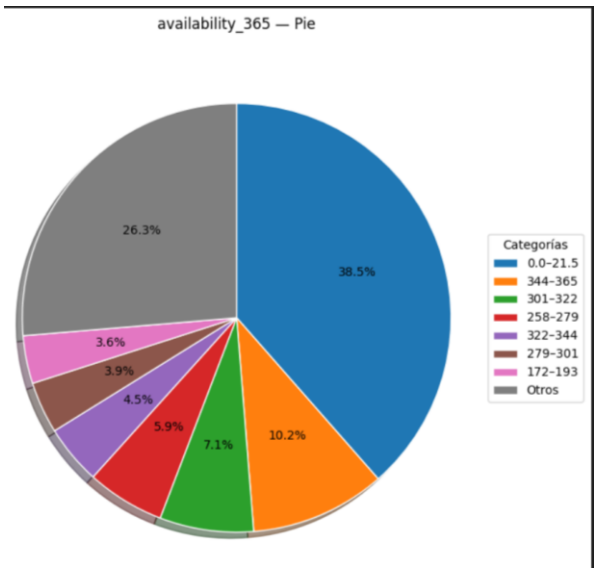
Interpretación: los precios están muy concentrados en el rango bajo–medio; el grupo “Otros” son outliers (listings muy caros o registros raros). Conviene auditarlos por posibles errores de captra o moneda.



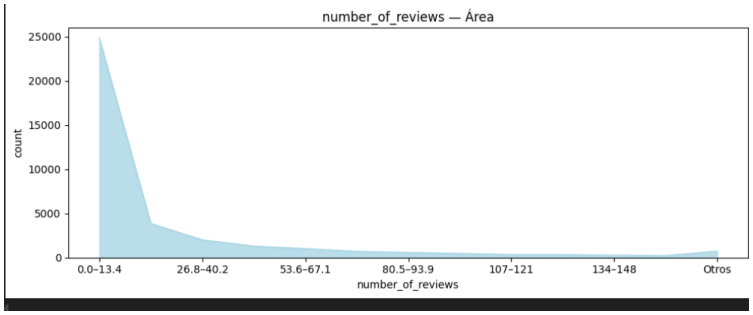
4) Qué es: agrupé availability_365 en rangos (Sturges) y armé un pastel para ver la proporción de días disponibles por anuncio.

Gráfico: domina 0–21.5 días con 38.5%; luego 344–365 con 10.2%, 301–322 con 7.1%, y franjas intermedias más pequeñas (5.9%, 4.5%, 3.9%, 3.6%). “Otros” suma 26.3%.

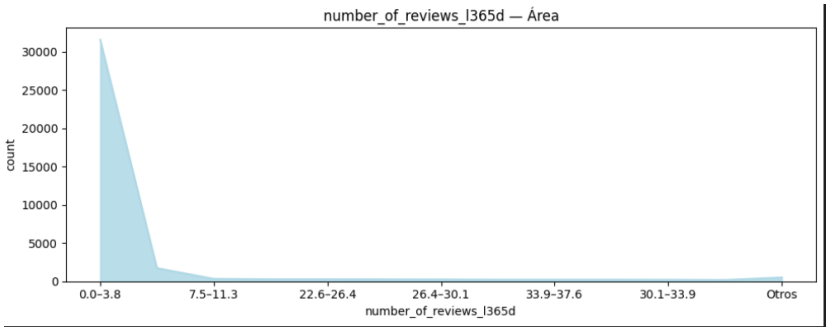
Interpretación: más de un tercio del inventario tiene muy poca disponibilidad (listings casi siempre ocupados o cerrados). Un 10% está casi todo el año disponible, posible oferta estructural o baja demanda. La distribución es bimodal; conviene segmentar precio/temporada para explicar estos dos polos.



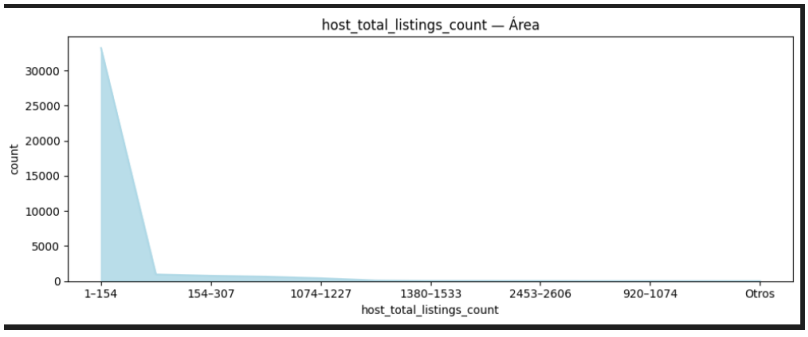
5) number_of_reviews — Área
Qué es: total de reseñas por anuncio, agrupadas en rangos (Sturges).
Gráfico: pico enorme en el primer tramo (~0–13) y luego caída rápida con cola larga.
Interpretación: la mayoría tiene pocas reseñas; unos pocos concentran muchas. Usaría mediana/percentiles y trataría aparte a los “top” por reputación.



6) number_of_reviews_l365d — Área
Qué es: reseñas solo del último año por anuncio, en rangos.
Gráfico: pico aún mayor en 0–4 y resto casi plano.
Interpretación: la actividad reciente es baja para la mayoría (muchos sin reseñas). Sirve para identificar listados inactivos/nuevos y enfocar campañas de activación.

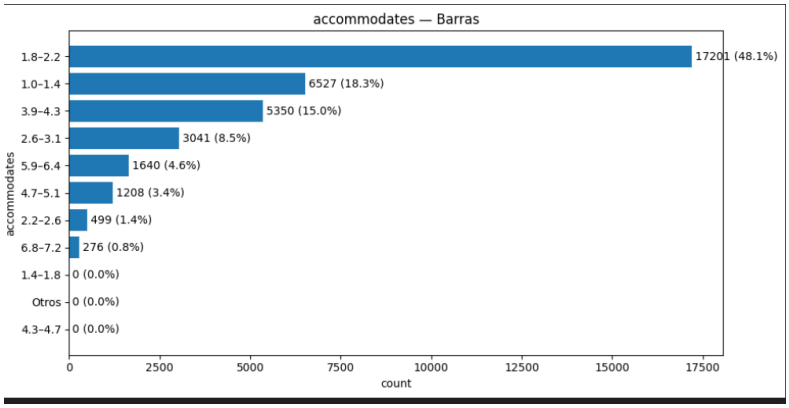


7) host_total_listings_count — Área
Qué es: cuántos anuncios tiene cada anfitrión en total.
Gráfico: casi todo cae en el primer tramo (1–154); luego la curva se aplana con una cola muy larga.
Interpretación: el mercado está dominado por anfitriones pequeños (1–3 anuncios); hay pocos "multi-propiedad" muy grandes. Para verlo mejor conviene re-agrupar en bins como 1, 2–5, 6–20, 21+ o usar escala log.

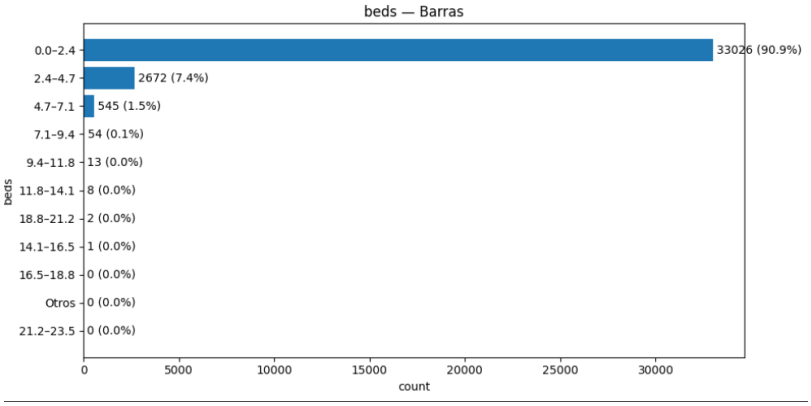


8) accommodates — Barras
Qué es: capacidad de huéspedes por anuncio (agrupada en rangos).
Gráfico: el tramo ~2 huéspedes lidera (~48%); le siguen ~1 huésped (~18%) y ~4 huéspedes (~15%); el resto es minoritario.

Interpretación: la oferta se concentra en estancias para 1–2 personas; unidades grandes son pocas. Útil para ajustar precios/paquetes y campañas orientadas a parejas o viajeros solos.

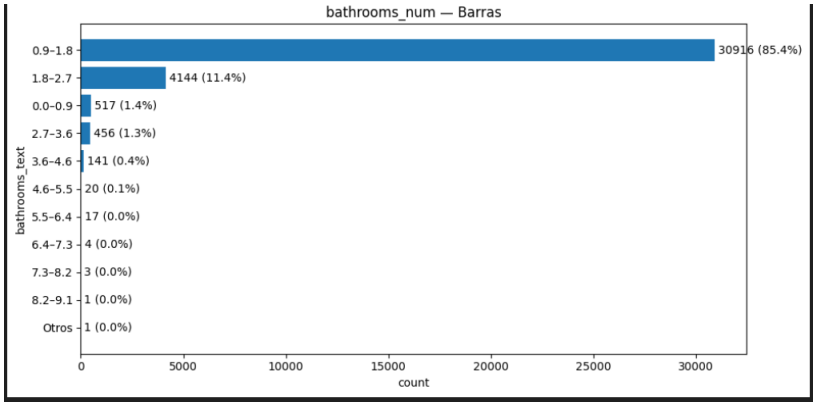


9) beds — Barras
Qué es: número de camas por anuncio (agrupado en rangos).
Gráfico: el rango 0.0–2.4 concentra 33,026 (90.9%); luego 2.4–4.7 con 2,672 (7.4%) y 4.7–7.1 con 545 (1.5%); el resto es marginal.
Interpretación: la oferta está ultra concentrada en 1–2 camas; listados con muchas camas son excepcionales.



10) bathrooms_num — Barras
Qué es: número de baños (numérico a partir de *bathrooms_text*), agrupado en rangos.
Gráfico: domina 0.9–1.8 con 30,916 (85.4%); después 1.8–2.7 con 4,144 (11.4%); hay pocos por debajo de 1 baño (517; 1.4%) y por encima de 2.7 (≤1.3% cada tramo).

Interpretación: la mayoría son 1 baño (algunos 1.5); 2 baños es minoría y más de 2 es raro—coincide con unidades pequeñas/medianas.



Conclusión

Viendo este corte de Airbnb en Nueva York, yo la percibo como un mercado enorme pero muy "compacto": la oferta se concentra en unidades pequeñas (1–2 camas y 1 baño) y en dos tipos claros, *entire rental unit* y *private room*, con precios metidos casi todos en un mismo rango y muy pocos outliers. Los calendarios están polarizados: hay muchísimos anuncios con 0 días disponibles en el año (anuncios pausados/bloqueados o fuera de temporada) y otro bloque con disponibilidad alta; eso deja la sensación de stock parcialmente inactivo. La demanda que reflejan las reseñas es de cola larga: la mayoría de listados tiene pocas, con unos cuantos muy grandes tirando el promedio. Del lado del anfitrión, predominan no-superhosts y hay ruido "unknown" en tasas y tiempos de respuesta, aunque la aceptación sí aparece mayormente alta, lo que sugiere que, cuando llega la solicitud, suelen aprobarla. Geográficamente se ve concentración por vecindarios (pareto marcado) y muchos registros "unknown", así que conviene limpiar y estandarizar esa columna. En resumen: Nueva York es masiva, atomizada en alojamientos pequeños, con fuerte concentración por zonas y un volumen no menor de anuncios inactivos; para actuar, yo priorizaría los barrios top, los *entire units* y una revisión de calidad de datos (response rate/time y location) antes de ajustar precios y disponibilidad.