Reporte - NYC Taxi Trip Duration (Reto 2)

Generado: 2026-01-05 20:14

Resumen ejecutivo

Este reporte muestra un flujo completo de ingeniería de datos con Pandas sobre el dataset NYC Taxi (~1.4M filas). Se aplican técnicas de optimización de memoria (downcasting y category), ingeniería de variables vectorizada (tiempo y geoespacial), visualización con storytelling y generación automática de un reporte PDF para toma de decisiones.

1) Optimización de memoria

Memoria antes: 417.32 MB

Memoria después: 134.93 MB

Reducción: 67.7%

Decisiones técnicas:
- Downcasting: int64 -> int8/int16/int32 según rangos reales.
- Float64 -> float32 para coordenadas y métricas continuas.
- Object -> category cuando la cardinalidad es baja (ej: store_and_fwd_flag).
- Fechas a datetime para habilitar .dt y mejorar vectorización.

2) Hallazgos clave (métricas)

- Hora pico: 18:00 (viajes: 90,600)
- Día con más viajes: Viernes
- Duración media: 15.99 min
- Distancia media: 3.44 km
- Velocidad media: 14.42 km/h
- Impacto del tráfico (pico vs valle): 37.9%
- Diferencia fin de semana: -4.8% (duración)

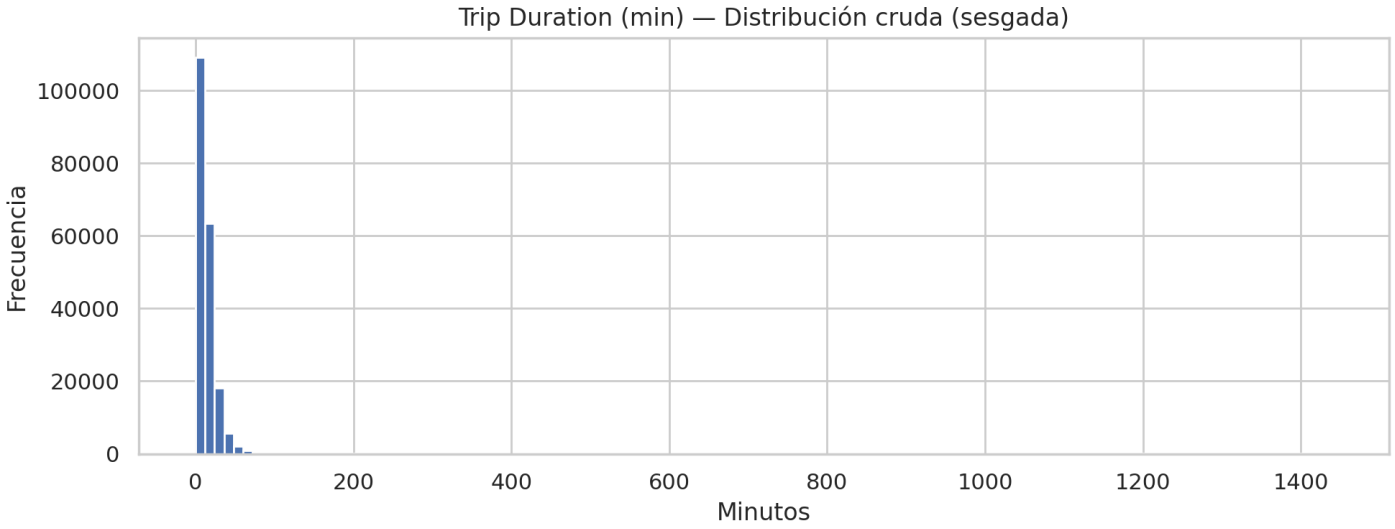
Interpretación rápida:
Las métricas combinan demanda (volumen), tiempo (duración), espacio (distancia) y eficiencia (velocidad). Esto permite contar una historia operacional: cuándo hay más viajes, dónde se concentran y cómo cambia la movilidad a lo largo del día.

Reporte - NYC Taxi Trip Duration (Reto 2)

Generado: 2026-01-05 20:14

3) Distribución de duración (cruda)

Esta gráfica suele verse extremadamente sesgada: la mayoría de viajes dura pocos minutos y existe una cola larga de viajes muy grandes (posibles outliers, errores de registro o casos raros). Sirve para justificar por qué una escala normal no permite ver bien el comportamiento típico.

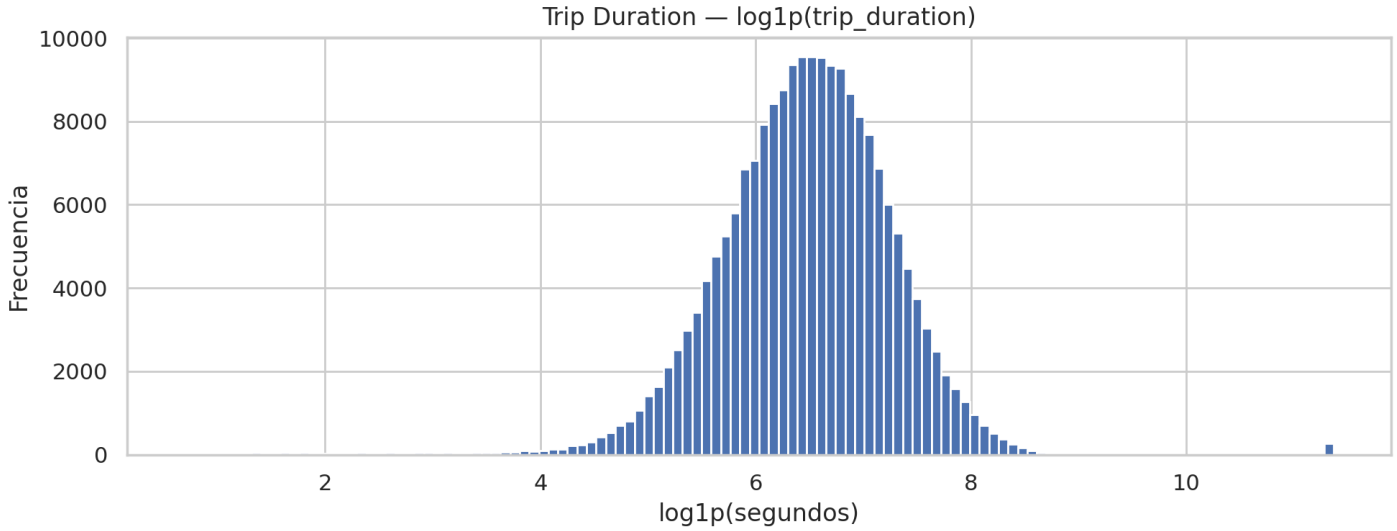


Reporte - NYC Taxi Trip Duration (Reto 2)

Generado: 2026-01-05 20:14

4) Distribución de duración (transformación log)

Aplicar log1p(duración) comprime la cola larga y revela mejor la forma central de la distribución. Esto facilita comparar patrones, detectar anomalías y construir modelos o reglas de negocio sin que los outliers dominen.

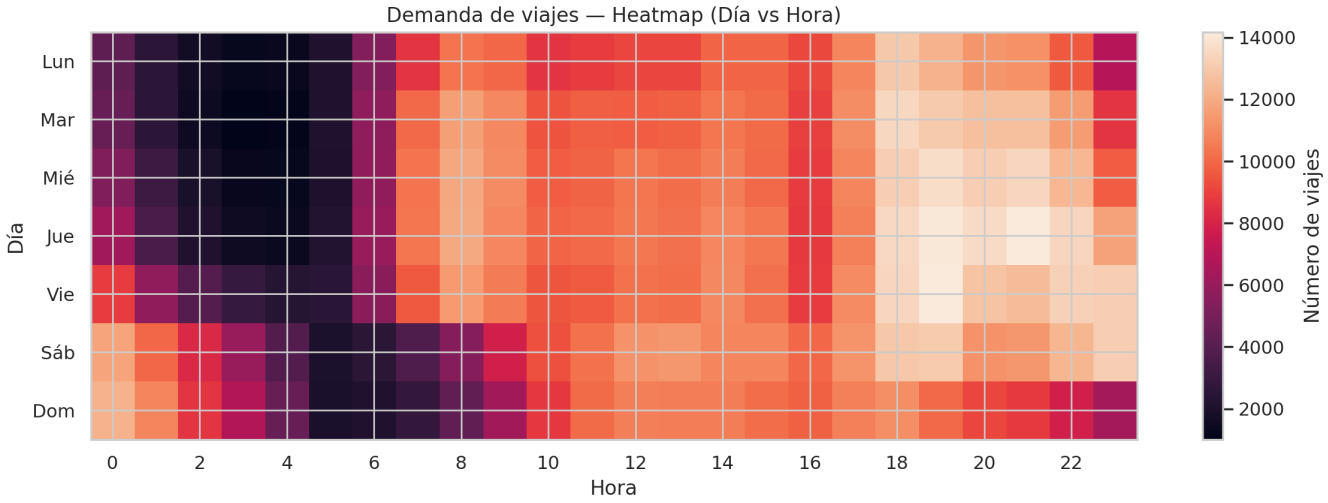


Reporte - NYC Taxi Trip Duration (Reto 2)

Generado: 2026-01-05 20:14

5) Demanda por día y hora (heatmap)

El heatmap muestra concentración de viajes por hora y día. Normalmente aparecen 'horas pico' (mañana y tarde) en días laborales, y patrones distintos en fin de semana. Es útil para decisiones de operación: turnos, oferta y planificación.

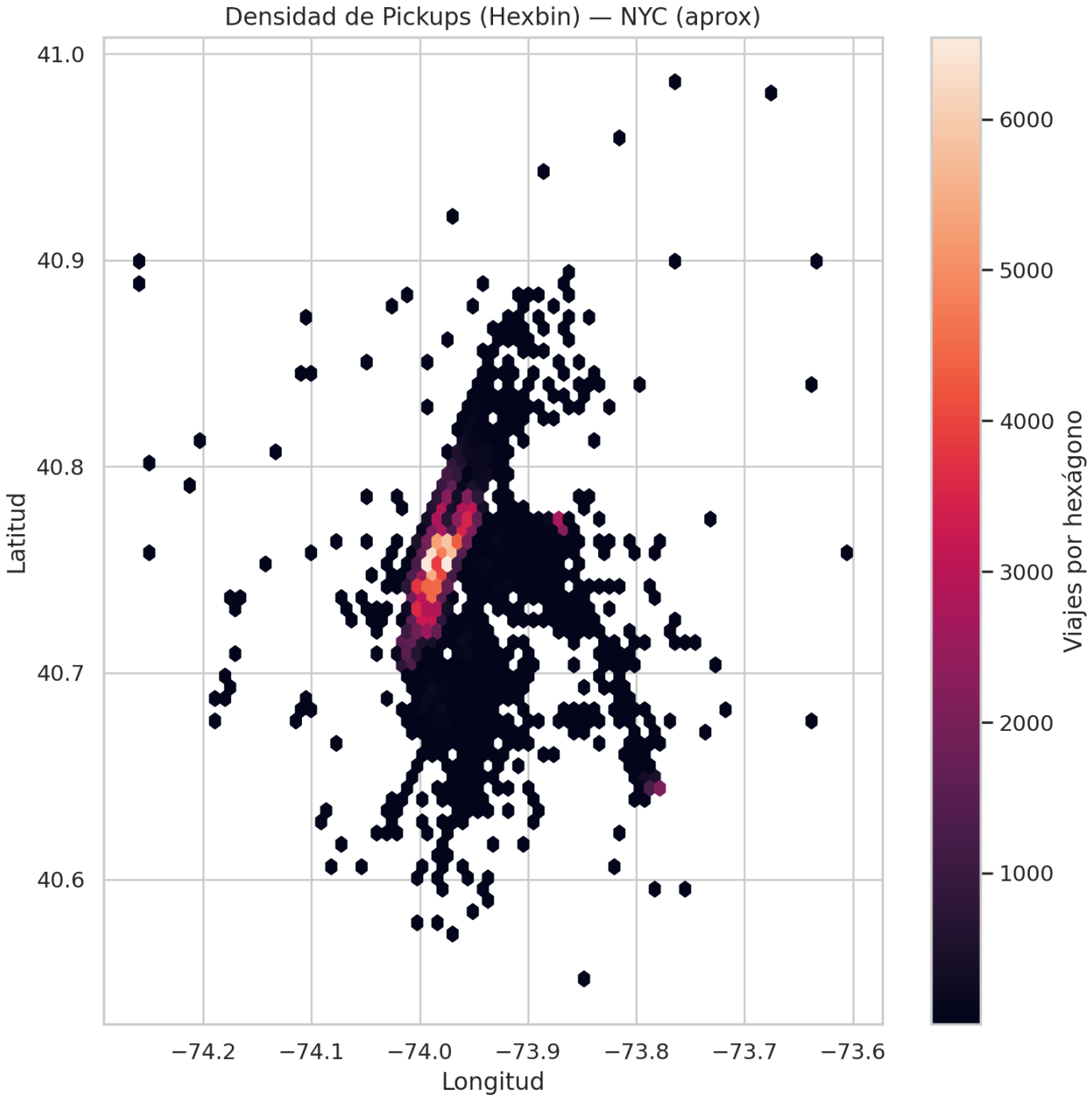


Reporte - NYC Taxi Trip Duration (Reto 2)

Generado: 2026-01-05 20:14

6) Densidad geoespacial (hexbin)

El hexbin resuelve el overplotting: con millones de puntos, un scatter se vuelve una mancha. Aquí cada hexágono agrega cantidad de pickups en esa zona. Zonas más claras = más demanda. Esto permite identificar hotspots (ej. Manhattan) y áreas de baja demanda.



Reporte - NYC Taxi Trip Duration (Reto 2)

Generado: 2026-01-05 20:14

7) Velocidad promedio por hora

Esta curva cuenta la historia del tráfico: en horas pico la velocidad cae (congestión), y en la madrugada sube (calles despejadas). Es una visual clave para medir eficiencia y estimar tiempos esperados.