Caso de estudio: Optimización del transporte en Nueva York:

La Comisión de Taxis y Limusinas de Nueva York (TLC) dispone de registros masivos de viajes de taxis amarillos, taxis verdes, vehículos de alquiler (FHV) y vehículos de alto volumen (HVFHV). Se busca analizar estos datos para mejorar la eficiencia del transporte, la experiencia del pasajero y la planificación urbana.

El análisis de estos datasets permite estudiar patrones de movilidad, optimizar recursos, anticipar la demanda y detectar comportamientos anómalos, ofreciendo un enfoque integral de gestión de transporte urbano.

Problemáticas que se pueden abordar con los datasets

1. **Predicción de forma de pago (Supervisado – Clasificación)**

Problema:

Muchos pasajeros pagan en efectivo y otros con tarjeta. Conocer de manera anticipada la forma de pago ayuda a planificar la logística del flujo de efectivo y la disponibilidad de métodos de cobro.

Objetivo:

Predecir si un pasajero pagará en efectivo o con tarjeta.

Mejorar la asignación de conductores a zonas según patrones de pago predominantes.

Optimizar la gestión de flujos de caja y reducir errores en cobros.

Dataset aplicables:

Yellow Taxi

Green Taxi

1. **Estimación de la duración y tarifa del viaje (Supervisado – Regresión)**

Problema:

Las tarifas y duraciones de los viajes pueden variar significativamente según la hora del día, la distancia y la zona. Los pasajeros buscan información precisa para planificar sus traslados.

Objetivo:

Predecir la tarifa total o la duración estimada de un viaje.

Mejorar la transparencia y confianza del pasajero.

Ayudar a la TLC a optimizar rutas y reducir tiempos de espera.

Dataset aplicable:

Yellow Taxi

Green Taxi

1. **Segmentación de zonas de la ciudad (No supervisado – Clustering)**

Problema:

Algunas zonas tienen alta concentración de viajes en ciertas horas, mientras que otras permanecen con poca actividad. Esto dificulta la distribución eficiente de taxis y vehículos de alquiler.

Objetivo:

Agrupar zonas de la ciudad según volumen de viajes, hora pico y tipo de servicio.

Identificar áreas con alta demanda o saturadas.

Optimizar la asignación de flotas y planificación de estaciones de vehículos.

Dataset aplicable:

Yellow Taxi

Green Taxi

FHV / HVFHV

1. **Detección de anomalías en viajes (No supervisado – Outlier detection)**

Problema:

Se detectan viajes con tarifas inusualmente altas, distancias inconsistentes o duraciones irreales, que podrían indicar errores de registro o intentos de fraude.

Objetivo:

Identificar viajes atípicos o fuera de patrón.

Mejorar la calidad y confiabilidad de los datos.

Minimizar pérdidas económicas y riesgos operativos.

Dataset aplicable:

Todos los datasets (Yellow, Green, FHV, HVFHV)

1. **b**

Problema:

Durante horas punta o eventos especiales, algunos tipos de vehículo pueden escasear en ciertas zonas, generando largas esperas y desbalance en la distribución de flotas.

Objetivo:

Predecir la demanda de cada tipo de vehículo según hora del día, día de la semana y zona.

Planificar anticipadamente la cantidad de vehículos necesarios en cada área.

Reducir tiempos de espera y mejorar la satisfacción de pasajeros.

Dataset aplicable:

Todos los datasets (Yellow, Green, FHV, HVFHV)

La columna service\_type será clave para diferenciar tipos de vehículo.

**1. FHV Trip Records (For-Hire Vehicles)**

Contiene viajes en vehículos de arriendo con chofer (distintos a taxis amarillos o verdes).

Variables clave:

dispatching\_base\_num: base que despachó el viaje.

pickup\_datetime y dropOff\_datetime: inicio y fin del viaje.

PUlocationID y DOlocationID: zonas TLC de inicio y fin.

SR\_Flag: si fue un viaje compartido (ej. Uber Pool, Lyft Line).

Affiliated\_base\_number: base con la que el vehículo está afiliado.

Importante: desde 2019 los viajes de alto volumen (Uber, Lyft, Juno, Via) se separan en otro dataset.

**2. Green Taxi Trip Records (LPEP)**

Describe viajes de los llamados taxis verdes o SHL (Street-Hail Livery).

Variables destacadas:

VendorID: empresa proveedora (ej. CMT, Curb, Myle).

lpep\_pickup\_datetime y lpep\_dropoff\_datetime: inicio y fin del viaje.

RatecodeID: tarifa aplicada (estándar, aeropuerto JFK/Newark, negociada, grupal, etc.).

passenger\_count, trip\_distance, fare\_amount, tip\_amount, tolls\_amount.

payment\_type: tipo de pago (tarjeta, efectivo, etc.).

trip\_type: viaje tomado en la calle o despachado por llamada.

Incluye cargos adicionales como congestion\_surcharge y desde 2025 el cbd\_congestion\_fee.

**3. High Volume FHV Trip Records (HVFHS)**

Viajes de empresas de alto volumen (más de 10.000 viajes diarios): Uber, Lyft, Via, Juno.

Variables clave:

hvfhs\_license\_num: licencia de la empresa (Uber, Lyft, etc.).

request\_datetime, pickup\_datetime, dropoff\_datetime.

trip\_miles, trip\_time.

Costos: base\_passenger\_fare, tolls, sales\_tax, congestion\_surcharge, airport\_fee.

tips, driver\_pay.

Flags:

shared\_request\_flag / shared\_match\_flag → si se solicitó o efectivamente se compartió el viaje.

access\_a\_ride\_flag → viajes gestionados por la MTA.

wav\_request\_flag y wav\_match\_flag → accesibilidad para silla de ruedas.

Incluye también el cbd\_congestion\_fee desde 2025.

**4. Yellow Taxi Trip Records (TPEP)**

Viajes de taxis amarillos tradicionales.

Variables principales:

VendorID: proveedor del sistema del taxímetro (CMT, Curb, etc.).

tpep\_pickup\_datetime y tpep\_dropoff\_datetime.

passenger\_count, trip\_distance, RatecodeID.

store\_and\_fwd\_flag: si los datos se almacenaron antes de ser enviados.

PULocationID, DOLocationID.

Tarifas: fare\_amount, extra, mta\_tax, tip\_amount, tolls\_amount, improvement\_surcharge, total\_amount.

payment\_type: forma de pago.

Recargos: congestion\_surcharge, airport\_fee, cbd\_congestion\_fee.

FHV → autos de arriendo con chofer, viajes individuales o compartidos.

LPEP (Green Taxi) → taxis verdes (street-hail livery).

HVFHS → Uber, Lyft, Via, Juno y similares (alto volumen).

Yellow Taxi → taxis amarillos tradicionales.

Datasets:

url\_1 = "https://d37ci6vzurychx.cloudfront.net/trip-data/yellow\_tripdata\_2025-06.parquet"

registros\_viajes = pd.read\_parquet(url\_1, engine="pyarrow")

url\_2 = "https://d37ci6vzurychx.cloudfront.net/trip-data/green\_tripdata\_2025-06.parquet"

registros\_viajes\_eco = pd.read\_parquet(url\_2, engine="pyarrow")

url\_3 = "https://d37ci6vzurychx.cloudfront.net/trip-data/fhv\_tripdata\_2025-06.parquet"

registro\_vehiculos\_alquiler = pd.read\_parquet(url\_3, engine="pyarrow")

url\_4 = "https://d37ci6vzurychx.cloudfront.net/trip-data/fhvhv\_tripdata\_2025-06.parquet"

Registros\_vehiculos\_gran\_volumen = pd.read\_parquet(url\_4, engine="pyarrow")