Metadatos

taxis.parquet

DATAVISION

Los datos del archivo *taxis.parquet* contienen la información correspondiente a los viajes realizados en la ciudad de Nueva York (dentro y fuera del área de Manhattan). Estos datos fueron extraídos de la base de datos de la **NYC Taxi & Limousine Commission**. La TLC es la agencia responsable de la licencia y regulación de los taxis con medallón (amarillos) de la ciudad de Nueva York, los vehículos de alquiler (livery comunitarios, autos negros y limusinas de lujo), las camionetas de transporte colectivo y los vehículos de para tránsito. Esta entidad pública la información de viajes de manera mensual. Se recopilo la información correspondiente a los **Yellow Taxi Trip Records** y los **Green Taxi Trip Records** para todos los meses en el periodo entre septiembre de 2023 y agosto de 2024.

Variable	Tipo de dato	Tipo de variable	Descripción	Valores posibles
VendorID	int	categórica	Código que indica el proveedor de TPEP que proporcionó el registro.	**1 =** Creative Mobile Technologies, LLC **2 =** VeriFone Inc.
tpep_pickup_datetime	datetime	fecha	La fecha y hora en que se activó el taxímetro.	
tpep_dropoff_datetime	datetime	fecha	La fecha y hora en que se desactivó el taxímetro.	
passenger_count	float	cuantitativa	La cantidad de pasajeros en el vehículo. Este es un valor ingresado por el conductor.	
trip_distance	float	cuantitativa	La distancia del viaje transcurrida en millas, reportada por el	

			taxímetro.	
RateCodeID	float	categórica	Código de tarifa final en vigor al final del viaje.	**1 =** Standard rate **2 =** JFK **3 =** Newark **4 =** Nassau or Westchester **5 =** Negotiated fare **6 =** Group ride
store_and_fwd_flag	string	categórica	Este indicador muestra si el registro del viaje se almacenó en la memoria del vehículo antes de enviarse al proveedor, conocido como “almacenar y reenviar,” debido a que el vehículo no tenía conexión con el servidor.	**Y =** Yes **N =** No
PULocationID	int	categórica	Zona de taxi TLC en la que se activó el taxímetro.	
DOLocationID	int	categórica	Zona de taxi TLC en la que se desactivó el taxímetro.	
payment_type	float	categórica	Código numérico que indica cómo pagó el pasajero el viaje.	**1 =** Credit card **2 =** Cash **3 =** No charge **4 =** Dispute **5 =** Unknown **6 =** Voided trip
fare_amount	float	cuantitativa	La tarifa por tiempo y distancia calculada por el taxímetro.	
extra	float	cuantitativa	Extras y recargos misceláneos. Actualmente, esto solo incluye los cargos de $0.50 y $1 por hora pico y horario nocturno.	
mta_tax	float	cuantitativa	Impuesto MTA de $0.50 que se activa automáticamente según la tarifa medida en uso.	

tip_amount	float	cuantitativa	Monto de la propina – Este campo se llena automáticamente para las propinas con tarjeta de crédito. Las propinas en efectivo no están incluidas.	
tolls_amount	float	cuantitativa	Monto total de todos los peajes pagados en el viaje.	
improvement_surcharge	float	cuantitativa	Recargo de mejora de $0.30 aplicado en los viajes al inicio de la carrera. Este recargo comenzó a aplicarse en 2015.	
total_amount	float	cuantitativa	El monto total cobrado a los pasajeros. No incluye propinas en efectivo.	
congestion_surcharge	float	cuantitativa	Monto total recaudado en el viaje por el recargo de congestión del estado de NY.	
airport_fee	float	cuantitativa	$1.25 solo para recogidas en los aeropuertos de LaGuardia y John F. Kennedy.	
type	string	categórica	Tipo de taxi en el que se realizó el viaje.	yellow, green
ehail_fee	float	-	Columna no se encuentra en los diccionarios encontrados en la pagina de TLC.	
trip_type	float	categórica	Código que indica si el viaje fue una parada en la calle o una solicitud por despacho, asignado automáticamente según la tarifa medida en uso, pero que el conductor puede modificar	**1 =** Street-hail **2 =** Dispatch

Consideraciones

La base de datos contiene 40.176.771 registros. En 8 de sus columnas se encuentran valores nulos y solo existen 4 duplicados para todas las columnas. En un análisis posterior se determinará que columnas serán empleadas para análisis posteriores, cuales serán transformadas o normalizadas y si existen columnas que serán eliminadas.

Variable	Observación	Acciones
VendorID	Se encontró en los datos una categoría adicional a las reportadas (6) para 3.475 registros.	Se evaluará si estos datos se consideraran como nulos o si se eliminaran los registros correspondientes.
tpep_pickup_datetime	Se encontró que se tienen 51 valores por fuera del periodo de análisis (2002, 2008, 2009 y 2026).	Se eliminarán estos registros.
tpep_dropoff_datetime	Se encontró que se tienen 52 valores por fuera del periodo de análisis (*1970, 2002, 2003, 2008, 2009 y 2026*)	Se eliminarán estos registros.
RatecodeID	Se encontrón en los datos una categoría adicional a las reportadas (99) para 384.161 registros.	Se evaluará si estos datos se consideraran como nulos o si se eliminaran los registros correspondientes.
store_and_fwd_flag	Esta es una variable categórica identificada con letras	Se realizará la transformación de ‘Y’ a 1 y de ‘N’ a 0, en caso de ser incluida en el modelo de Machine Learning.
payment_type	Las categorías se encontraban enumeradas del 0 al 5, en lugar de del 1 al 6.	Se realizará el cambio a los valores establecidos en el diccionario de datos.
type	Esta es una variable categórica identificada con palabras.	Se realizará la transformación de ‘Yellow’ a 1 y de ‘Green’ a 2, en caso de ser incluida en el modelo de Machine Learning.

ehail_fee	Esta columna contiene solo nulos y no se encuentra incluida en el diccionario de datos oficial de la fuente.	Se procederá a eliminar esta columna del DataFrame.
Variables cuantitativas	Se encontraron valores negativos para variables que representan distancias y tarifas.	Se evaluará si estos datos serán eliminados, imputados, se aplicará valor absoluto o se cambiarán a cero.