Universidad Nacional de San Agustín de Arequipa
Escuela Profesional de Ciencia de la Computación
Curso: Tópicos en Ciencia de Datos

Práctica de Laboratorio: Análisis Exploratorio de Datos - Data Wrangling

Profesora: Ana Maria Cuadros Valdivia
Alumno: Saúl Arturo Condori Machaca

1. Contexto

El presente estudio se enmarca dentro del análisis de movilidad urbana y la gestión de riesgos ante eventos sísmicos en la ciudad de Nueva York. Con el objetivo de identificar zonas críticas y rutas potencialmente vulnerables en caso de evacuación, se emplean dos fuentes de datos complementarias que permiten el análisis temporal y geoespacial de la actividad urbana y sísmica.

1.1. Datos de trayectos de taxis en NYC

El primer conjunto de datos corresponde al **NYC Yellow Taxi Trip Data**, provisto por la *New York City Taxi & Limousine Commission (TLC)*. Este dataset contiene registros detallados de viajes realizados por taxis amarillos, los cuales representan un importante medio de transporte urbano en la ciudad.

Entidad u objeto de estudio: Cada registro en este conjunto de datos representa un *viaje individual de taxi*, caracterizado por su ubicación geográfica, duración, distancia, número de pasajeros y forma de pago.

Este conjunto de datos contiene un total de **11 382 049 registros**, lo que refleja una cobertura significativa del comportamiento de movilidad urbana en la ciudad.

A continuación se describen los atributos principales:

Cuadro 1: Descripción de los atributos del dataset *NYC Yellow Taxi Trip Data*

Atributo	**Descripción detallada**
VendorID	Código del proveedor de tecnología que generó el registro del viaje. Los valores posibles corresponden a empresas que manejan los taxímetros: *1 = Creative Mobile Technologies (CMT)*, *2 = VeriFone Inc.*
tpep_pickup_datetime	Fecha y hora exactas en que el viaje comenzó, es decir, cuando el pasajero fue recogido y el taxímetro activado.
tpep_dropoff_datetime	Fecha y hora exactas en que el viaje terminó, es decir, cuando el pasajero fue dejado y el taxímetro detenido.
passenger_count	Número de pasajeros que fueron transportados durante el viaje. Es ingresado manualmente por el conductor.
trip_distance	Distancia total recorrida durante el viaje, medida en millas.
pickup_longitude / pickup_latitude	Coordenadas geográficas (longitud/latitud) del punto de inicio del viaje.
dropoff_longitude / dropoff_latitude	Coordenadas geográficas (longitud/latitud) del punto final del viaje.
RatecodeID	Código de tarifa aplicada al viaje. Incluye valores como tarifa estándar, tarifa hacia aeropuertos como JFK o Newark, tarifas negociadas, etc.
store_and_fwd_flag	Indica si el registro fue almacenado temporalmente en el vehículo antes de ser enviado al servidor por falta de conexión (Y = *sí*, N = *no*).
payment_type	Código numérico que representa el tipo de pago utilizado: 1 = tarjeta de crédito, 2 = efectivo, 3 = sin cargo, 4 = disputa, 5 = desconocido, 6 = viaje anulado.
fare_amount	Monto base cobrado por el viaje, calculado según tiempo y distancia recorrida.
extra	Cargos adicionales como el recargo nocturno o por hora pico.
mta_tax	Impuesto obligatorio de $0.50 destinado a la Autoridad Metropolitana de Transporte.
tip_amount	Monto de propina recibido. Solo se registra si se paga con tarjeta; las propinas en efectivo no están registradas.
tolls_amount	Monto total pagado por peajes durante el viaje.
improvement_surcharge	Recargo de $0.30 aplicado desde 2015 para mejoras en el servicio de taxis.
total_amount	Monto total cobrado al pasajero, incluyendo tarifa base, extras, impuestos, propinas y peajes.

Cuadro 2: Resumen de atributos del dataset de taxis

Atributo	**Tipo de dato**	**Mínimo**	**Máximo**
VendorID	int64	1	2
tpep_pickup_datetime	object	2016-02-01	2016-02-29
tpep_dropoff_datetime	object	2015-02-07	2016-06-26
passenger_count	int64	0	9
trip_distance	float64	-3,390,583.8	11,658,534.3
pickup_longitude	float64	-130.82	94.64
pickup_latitude	float64	-77.03	59.35
RatecodeID	int64	1	99
store_and_fwd_flag	object	'N'	'Y'
dropoff_longitude	float64	-122.61	38.90
dropoff_latitude	float64	-77.03	405.32
payment_type	int64	1	4
fare_amount	float64	-450.0	154,810.43
extra	float64	-47.6	637.97
mta_tax	float64	-1.0	80.5
tip_amount	float64	-35.0	622.11
tolls_amount	float64	-99.99	913.0
improvement_surcharge	float64	-0.3	0.3
total_amount	float64	-450.3	154,832.14

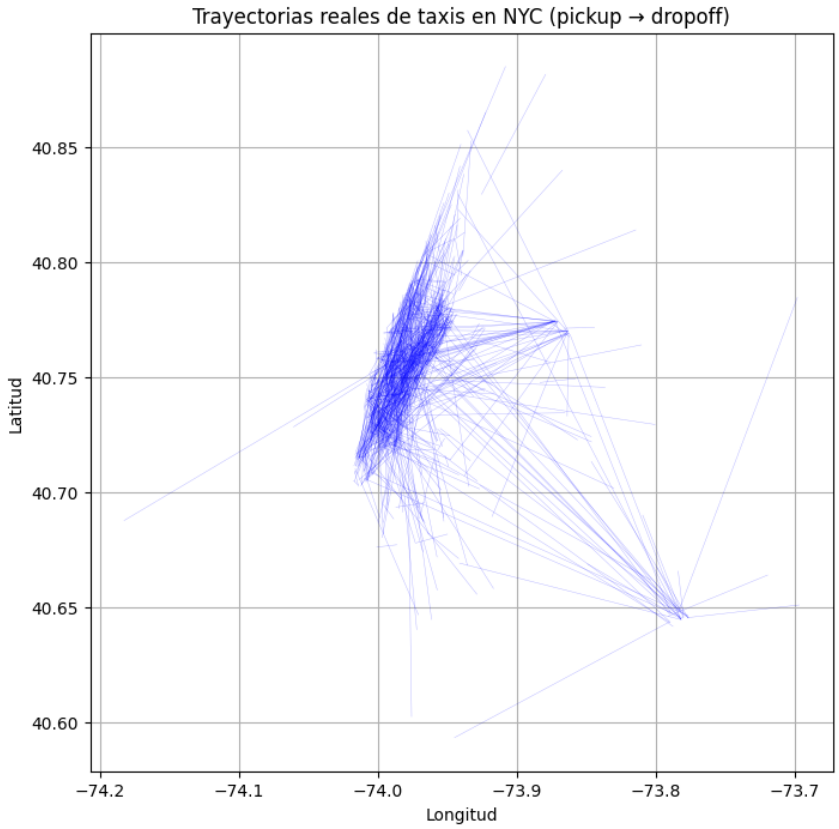

Figura 1: Distribución geográfica de las trayectorias de algnuos taxis amarillos en NYC. Cada linea representa una trayectoria.

1.2. Datos de actividad sísmica

El segundo conjunto de datos utilizado es el **Earthquakes Data NY**, extraído del *Servicio Geológico de los Estados Unidos (USGS)*. Este dataset contiene información de eventos sísmicos registrados en la región noreste de los Estados Unidos, incluyendo Nueva York y áreas colindantes.

Entidad u objeto de estudio: Cada registro en este conjunto representa un *evento sísmico individual*, con información sobre su localización, magnitud, profundidad y características del fenómeno.

Este conjunto de datos contiene un total de **1 203 registros**, representando eventos ocurridos en fechas recientes, tanto naturales (terremotos) como antrópicos (explosiones de cantera).

A continuación se describen los atributos más importantes:

Cuadro 3: Descripción de los atributos del dataset *Earthquakes Data NY*

Atributo	**Descripción detallada**
time	Fecha y hora del evento sísmico, en formato ISO (UTC).
latitude / longitude	Coordenadas geográficas (latitud y longitud) del epicentro del sismo.
depth	Profundidad del evento sísmico, medida en kilómetros.
mag	Magnitud del evento sísmico, valor numérico que refleja la energía liberada.
magType	Tipo de magnitud utilizada (ej. *ml* = magnitud local, *mb_lg* = magnitud de onda larga).
nst	Número de estaciones sísmicas que detectaron el evento.
gap	Ángulo máximo entre estaciones adyacentes, en grados. Un valor menor indica mejor cobertura.
dmin	Distancia mínima desde el epicentro a la estación más cercana, en grados.
rms	Raíz cuadrática media del ajuste entre los datos y el modelo sísmico. Mide la calidad del ajuste.
net	Código de red de monitoreo sísmico que registró el evento (ej. *us* = red USGS).
id	Identificador único del evento sísmico.
updated	Fecha y hora en que se actualizó por última vez la información del evento.
place	Descripción textual del lugar más cercano al epicentro (por ejemplo: "5 km W of Bedminster, NJ").
type	Tipo de evento: puede ser un *earthquake* (terremoto), *quarry blast* (explosión de cantera), etc.
horizontalError	Error estimado en la ubicación horizontal del epicentro, en kilómetros.
depthError	Error estimado en la medición de profundidad, en kilómetros.
magError	Error estimado de la magnitud del evento.
magNst	Número de estaciones que contribuyeron específicamente al cálculo de la magnitud.
status	Estado de revisión del evento: *reviewed* (revisado) o *automatic* (generado automáticamente).
locationSource	Código de la fuente responsable del cálculo de la ubicación del evento.
magSource	Código de la fuente responsable del cálculo de la magnitud.

Cuadro 4: Resumen de atributos del dataset de sismos

Atributo	**Tipo de dato**	**Mínimo**	**Máximo**
time	object	1929-08-12	2025-05-29
latitude	float64	40.66	45.18
longitude	float64	-79.87	-71.65
depth	float64	0.0	24.98
mag	float64	0.06	5.74
magType	object	'mb', 'ml', etc.	'mb', 'ml', etc.
nst	float64	3.0	245.0
gap	float64	27.0	341.0
dmin	float64	0.001	1.137
rms	float64	0.01	1.18
net	object	'us', 'se', etc.	'us', 'se', etc.
id	object	ID de evento sísmico	ID de evento sísmico
updated	object	Fecha ISO	Fecha ISO
place	object	Descripción textual	Descripción textual
type	object	'earthquake', etc.	'earthquake', etc.
horizontalError	float64	0.12	23.81
depthError	float64	0.22	42.6
magError	float64	0.02	0.65
magNst	float64	0.0	130.0
status	object	'automatic', etc.	'reviewed', etc.
locationSource	object	Código de fuente	Código de fuente
magSource	object	Código de fuente	Código de fuente

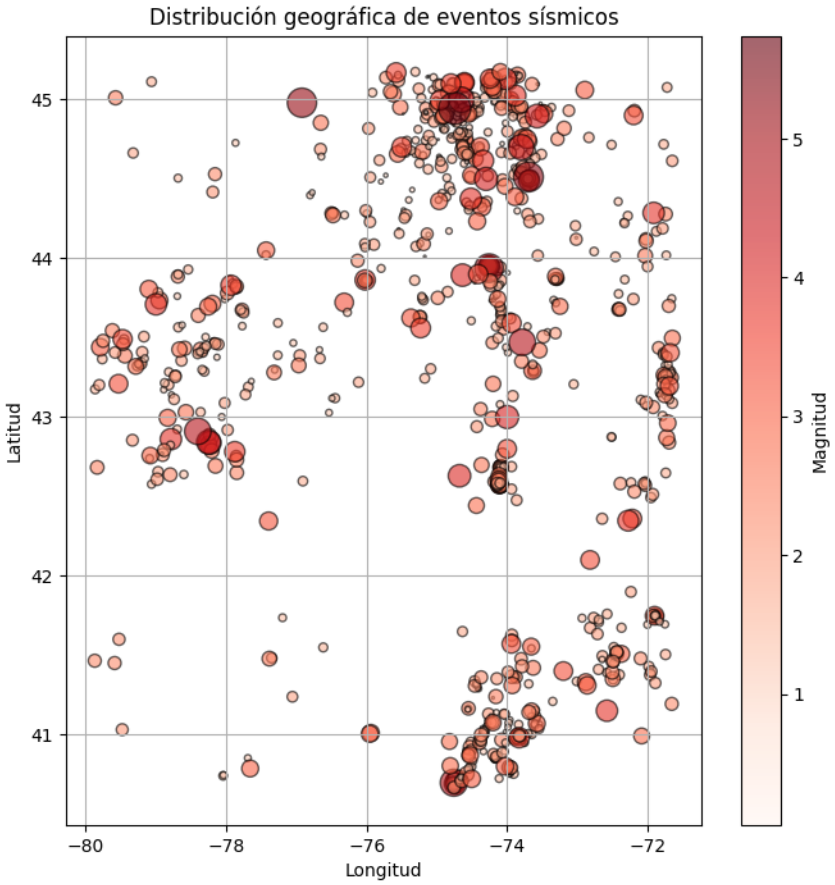

Figura 2: Distribución geográfica de los eventos sísmicos registrados. El tamaño y el color indican la magnitud.

2. Eliminación de Valores Nulos y Duplicados

Para garantizar la integridad de los datos y evitar sesgos en los análisis posteriores, se llevó a cabo un proceso de detección y limpieza de valores nulos y registros duplicados en ambos datasets utilizados en el estudio.

2.1. Dataset de Taxis

El análisis del conjunto de datos *NYC Yellow Taxi Trip Data* reveló que no existen valores nulos en ninguna de sus columnas. En cuanto a duplicados, se detectó un único registro repetido entre un total de más de 11 millones de observaciones. Aunque su impacto estadístico es insignificante, se optó por eliminarlo para mantener la consistencia estructural del dataset.

2.2. Dataset de Sismos

En el conjunto de datos *Earthquakes Data NY*, se identificaron valores nulos en varias columnas, como se resume a continuación:

- Las columnas clave para el análisis – `time`, `latitude`, `longitude`, `mag` y `type` – no presentan valores nulos.

- La variable `depth` presenta 10 registros con valores faltantes. Dado que la profundidad del sismo puede influir en su impacto urbano, se decidió eliminar estos registros.
- Otras variables con valores nulos (`magError`, `nst`, `gap`, etc.) no son esenciales para el análisis del presente estudio, por lo que se conservaron sin alteración.

No se identificaron registros duplicados en este dataset, por lo que no fue necesaria ninguna acción adicional en ese aspecto.

2.3. Resumen de la limpieza realizada

- **Taxis:** Se eliminó 1 registro duplicado.
- **Sismos:** Se eliminaron 10 registros con valor nulo en la variable `depth`.

Estas acciones aseguran que los datos utilizados en los análisis posteriores tengan coherencia estructural y relevancia temática de acuerdo con el objetivo del estudio: analizar trayectorias urbanas y eventos sísmicos para detectar zonas críticas y rutas vulnerables en escenarios de evacuación.

3. Identificación de Outliers y Posibles Anomalías

Durante el análisis estadístico preliminar mediante el método `describe()`, se identificaron posibles valores atípicos y anomalías en ambos conjuntos de datos. A continuación se presentan las observaciones detectadas y las acciones sugeridas para su tratamiento.

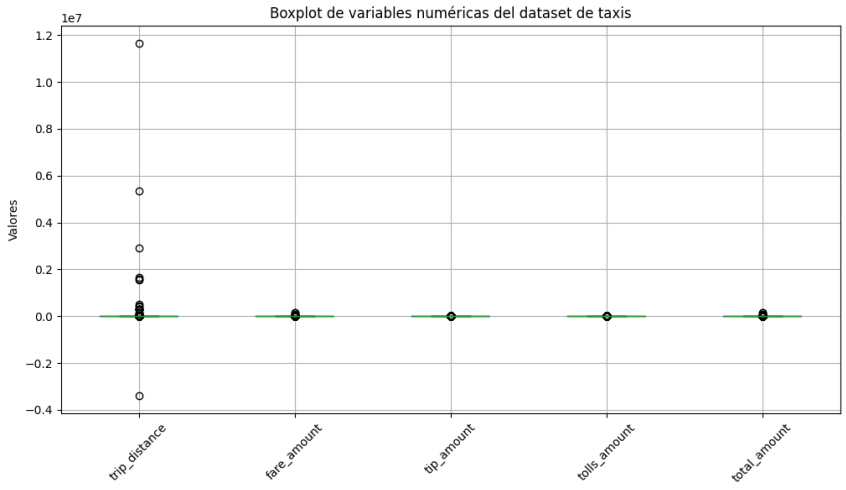

Figura 3: Identificacion de Valores Atípicos de la variable trip-distance, mediante el grafico BoxPlot.

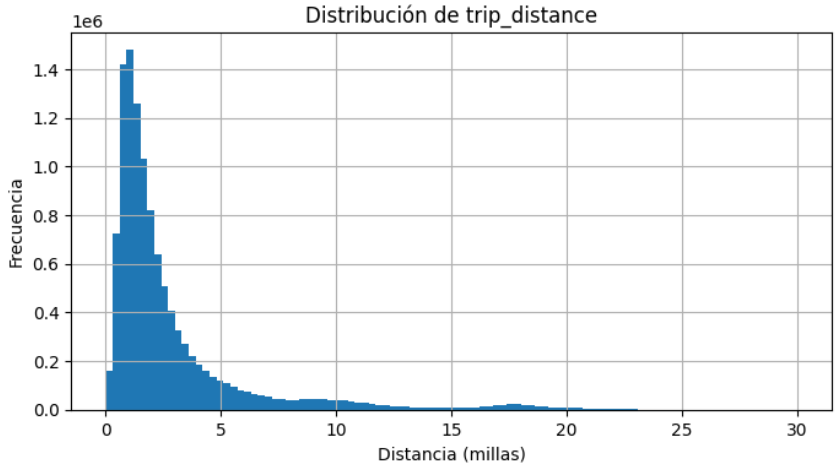

Figura 4: Identificacion de Valores Atípicos de la variable trip-distance, mediante un diagrama de frecuencias.

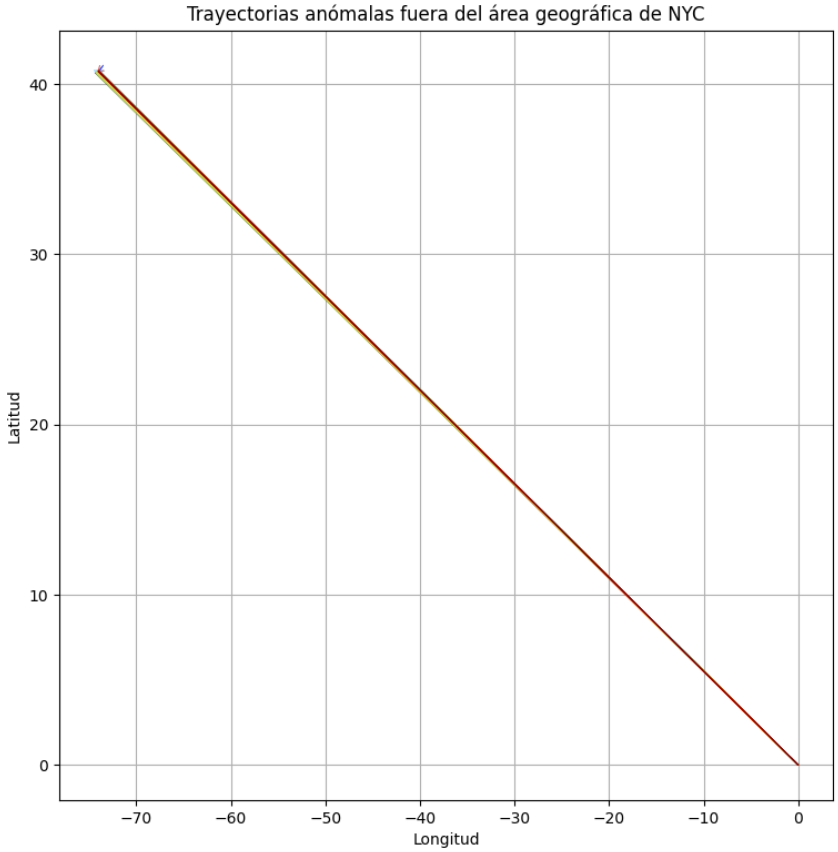

Figura 5: Identificacion de las trayectorias fuera del rango Geografico real de Nueva York.

Figura 6: Identificacion de las trayectorias fuera del rango Geografico real de Nueva York.

3.1. Dataset de Taxis (`taxis_df`)

Cuadro 5: Detección de outliers en el dataset *NYC Yellow Taxi Trip Data*

Variable	**Observación**	**Acción a tomar**
`trip_distance`	Mínimo: -3,390,584 Máximo: 11,658,530	Valores imposibles: distancias negativas o excesivamente altas. Filtrar a un rango razonable (por ejemplo, `0 <trip_distance <100`).
`pickup_longitude / pickup_latitude` `dropoff_longitude / dropoff_latitude`	Longitudes hasta 94 y latitudes hasta 405	Fuera del rango geográfico real de la ciudad de Nueva York. NYC está aproximadamente entre latitudes 40.4–41.0 y longitudes -74.3 a -73.6. Se recomienda filtrar los registros fuera de este rango.
`fare_amount / total_amount`	Presencia de valores negativos (hasta -450 dólares)	No es válido que las tarifas sean negativas. Se deben eliminar todos los registros con montos negativos.
`extra, mta_tax, tolls_amount, improvement_surcharge, tip_amount`	Valores negativos y valores extremadamente altos (por ejemplo, propinas mayores a 600 dólares)	Eliminar valores negativos. Considerar establecer un límite superior razonable para evitar que los valores extremos distorsionen los análisis.

3.2. Dataset de Sismos (`earthquakes_df`)

Cuadro 6: Evaluación de outliers en el dataset *Earthquakes Data NY*

Variable	Observación	Acción sugerida
`depth`	Mínimo: 0 km Máximo: 24.98 km	Rango aceptable para el contexto de análisis urbano. No se requiere transformación.
`mag`	Mínimo: 0.06 Máximo: 5.74	Las magnitudes menores a 1 corresponden a eventos sísmicos leves, y son frecuentes en áreas urbanas. No se requiere intervención.
`gap, rms, magError, horizontalError`	Presencia de valores altos en algunas métricas auxiliares	Estos valores reflejan la calidad del evento registrado. Aunque altos, están dentro del rango esperable en datos sísmicos. Se pueden conservar o analizar más adelante según los requerimientos del modelo.

4. Transformación y Limpieza de Outliers

Una vez identificadas las anomalías y valores atípicos en el dataset *NYC Yellow Taxi Trip Data*, se procedió a realizar una limpieza estructurada y justificada de los datos con el fin de asegurar la validez y representatividad de los análisis posteriores. Las transformaciones se realizaron conforme a los siguientes criterios:

4.1. Criterios de filtrado aplicados

- **Filtro geográfico:** Se conservaron únicamente los registros cuyas coordenadas de recogida y destino estuvieran dentro del rango geográfico real de la ciudad de Nueva York:
  - Latitud entre 40.4 y 41.0
  - Longitud entre -74.3 y -73.6
- **Filtro por distancia:** Se eliminaron los viajes con distancia igual o menor a cero, así como aquellos con valores extremos mayores a 100 millas, considerados como improbables en el contexto urbano.
- **Filtro por tarifas:** Se descartaron registros con valores negativos o nulos en las variables `fare_amount` y `total_amount`, ya que no representan viajes válidos.
- **Filtro por variables monetarias auxiliares:** Se eliminaron registros con valores negativos en variables como: `extra`, `mta_tax`, `tolls_amount`, `improvement_surcharge` y `tip_amount`, ya que conceptualmente estos no pueden ser menores que cero.

4.2. Resultados de la limpieza

Luego de aplicar los filtros mencionados, se obtuvieron los siguientes resultados:

- **Registros originales:** 11 382 048
- **Registros después de la limpieza:** 11 136 675
- **Total de registros eliminados:** 245 373

Esta limpieza permitió eliminar datos inconsistentes o irreales que podrían haber afectado negativamente la interpretación de trayectorias, la segmentación horaria o la modelización de rutas vulnerables. El dataset resultante posee coherencia estructural y es representativo del comportamiento real del sistema de transporte urbano en Nueva York.

5. Estandarización del Formato Temporal

Uno de los objetivos clave de este estudio es la segmentación de datos por franjas horarias (mañana, tarde y noche), lo cual requiere que las variables temporales en ambos datasets estén correctamente estructuradas y unificadas bajo un mismo formato.

5.1. Diagnóstico inicial

A continuación se muestra el tipo de dato inicial y un ejemplo representativo para la columna de tiempo en cada dataset:

Cuadro 7: Diagnóstico inicial del formato temporal en los datasets

Dataset	**Columna temporal**	**Tipo de dato inicial**	**Ejemplo de valor**
`taxis_df`	`tpep_pickup_datetime`	`object`	`2016-02-25 17:24:20`
`earthquakes_df`	`time`	`object`	`2025-05-29 14:30:05.220000+00:00`

En el caso de `taxis_df`, la columna de tiempo está en un formato legible pero sin especificación de zona horaria (conocido como *naive datetime*). Por otro lado, el dataset `earthquakes_df` almacena sus fechas en formato ISO 8601, con zona horaria UTC explícita.

5.2. Justificación de la estandarización

Dado que el análisis requiere comparar y clasificar eventos según la hora del día, fue necesario transformar ambas columnas al tipo de dato `datetime64[ns]`, sin zona horaria, para que sean comparables en términos absolutos de tiempo. Esta conversión asegura que se pueda aplicar lógica de franja horaria de manera coherente entre registros de movilidad urbana y eventos sísmicos.

5.3. Resultado de la conversión

Luego de aplicar la estandarización temporal, ambas columnas fueron convertidas exitosamente a un tipo común de fecha-hora sin zona horaria. En la siguiente tabla se presentan los tipos de datos resultantes y ejemplos de valores:

Cuadro 8: Resultado final de la conversión del formato temporal

Dataset	**Columna temporal**	**Tipo final**	**Ejemplo de valor post-conversión**

`taxis_df`	`tpep_pickup_datetime`	`datetime64[ns]`	`2016-02-25 17:24:20`
			`2016-02-25 23:10:50`
			`2016-02-01 00:00:01`

`earthquakes_df`	`time`	`datetime64[ns]`	`2025-05-29 14:30:05.220`
			`2025-05-20 18:30:05.998`
			`2025-05-19 17:29:14.565`

Con esta estandarización, ambos datasets están ahora alineados temporalmente y listos para ser analizados por franjas horarias de manera consistente.

6. Clasificación Temporal por Franjas Horarias

Con el objetivo de analizar el comportamiento urbano y la ocurrencia de eventos sísmicos en función del momento del día, se clasificaron todos los registros de ambos conjuntos de datos en franjas horarias. Esta segmentación temporal es esencial para identificar patrones diferenciados en la movilidad y evaluar la vulnerabilidad ante emergencias según la hora en que ocurren los eventos.

6.1. Definición de franjas horarias

Se definieron cuatro franjas horarias principales, basadas en divisiones convencionales del día:

- **Madrugada:** 00:00 a 05:59
- **Mañana:** 06:00 a 11:59
- **Tarde:** 12:00 a 17:59
- **Noche:** 18:00 a 23:59

Esta clasificación se aplicó a la columna `tpep_pickup_datetime` del dataset `taxis_df` y a la columna `time` del dataset `earthquakes_df`, generando una nueva columna en cada uno denominada `franja_horaria`.

6.2. Ejemplos de clasificación

A continuación se presentan ejemplos de cómo quedó la columna `franja_horaria` después de aplicar la transformación:

Cuadro 9: Ejemplos de franja horaria en el dataset de taxis

tpep_pickup_datetime	**franja_horaria**
2016-02-25 17:24:20	tarde
2016-02-25 23:10:50	noche
2016-02-01 00:00:01	madrugada
2016-02-01 00:05:16	madrugada
2016-02-01 00:20:59	madrugada

Cuadro 10: Ejemplos de franja horaria en el dataset de sismos

time	**franja_horaria**
2025-05-29 14:30:05	tarde
2025-05-20 18:30:05	noche
2025-05-19 17:29:14	tarde
2025-04-27 16:27:56	tarde
2025-04-25 00:17:58	madrugada

La incorporación de esta nueva variable categórica temporal permitirá realizar análisis desagregados por franjas del día, lo cual es especialmente útil para evaluar riesgos diferenciales en horas pico, baja movilidad o condiciones nocturnas.

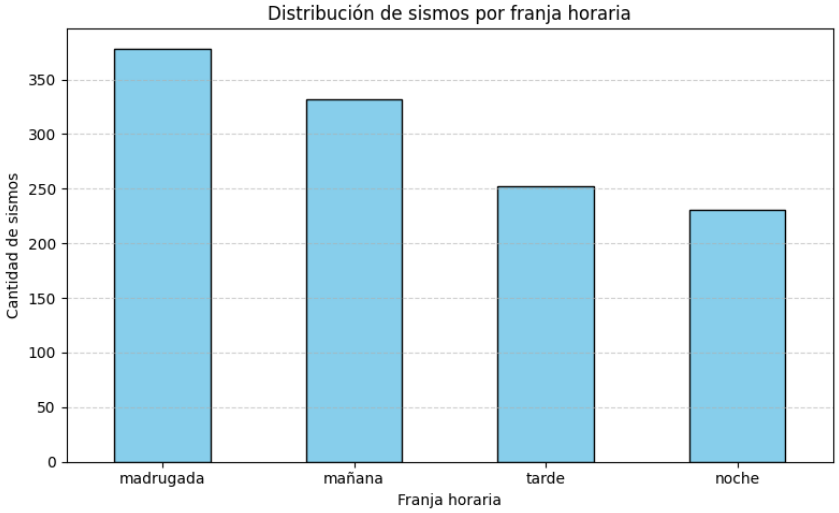

Figura 7: Gráfico de barras para distribución de sismos por franja horaria

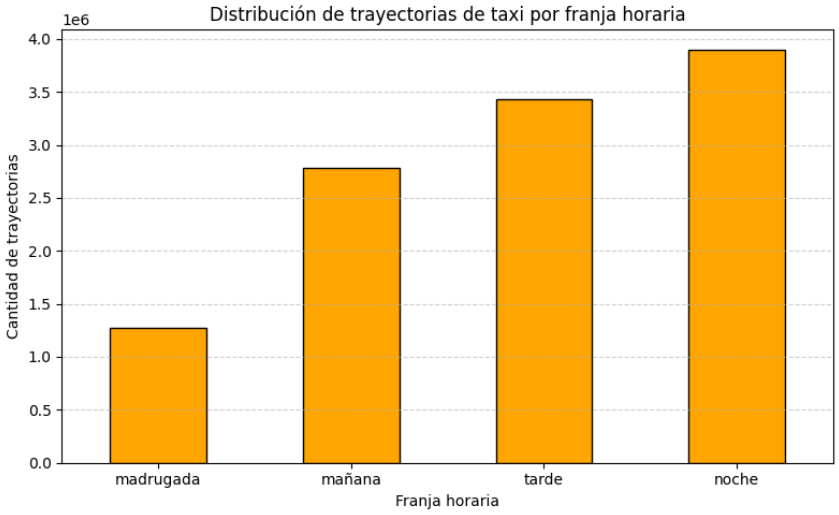

Figura 8: Gráfico de barras para distribución de trayectorias de taxi por franja horaria

7. Conclusiones

El presente análisis y transformación permitió integrar dos fuentes de datos de naturaleza distinta —movilidad urbana y actividad sísmica— con el objetivo de identificar posibles zonas críticas y evaluar riesgos en entornos urbanos densamente poblados como la ciudad de Nueva York.

A partir del proceso de limpieza, transformación y estandarización, se obtuvieron conjuntos de datos coherentes, confiables y comparables, tanto en términos espaciales como temporales. La eliminación de valores atípicos y registros inconsistentes mejoró sustancialmente la calidad analítica del dataset de taxis, mientras que la estandarización horaria permitió clasificar adecuadamente los eventos según franjas del día.

Entre los hallazgos más relevantes, destacan:

- La alta ocurrencia de trayectos de taxi durante la mañana y la tarde, coincidiendo con las horas de mayor actividad urbana.
- La ocurrencia de eventos sísmicos distribuidos a lo largo del día, con un número no despreciable durante la madrugada, lo cual podría representar un riesgo importante dada la baja capacidad de respuesta en esa franja horaria.
- La existencia de rutas de taxi que atraviesan zonas potencialmente vulnerables, lo que sugiere la necesidad de evaluar estrategias de evacuación considerando tanto la densidad de tránsito como la exposición sísmica.

Este estudio representa un primer paso hacia la creación de modelos visuales y predictivos que integren movilidad urbana y fenómenos naturales, contribuyendo al diseño de planes de respuesta ante desastres más eficientes y contextualizados.

Los datos y herramientas utilizados sientan las bases para futuras investigaciones en visual analytics y planificación urbana resiliente, siendo posibles líneas futuras de trabajo la incorporación de datos meteorológicos, de infraestructura crítica y simulaciones de evacuación en tiempo real.