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

Pipeline de Cs. de Datos

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

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

2. ¿Qué problemas identificas en el dataset?

Durante el análisis exploratorio inicial de los datasets utilizados en este estudio —*NYC Yellow Taxi Trip Data* y *Earthquakes Data NY*— se identificaron diversos problemas que comprometían la calidad y la representatividad de los datos, siendo necesarios procesos específicos de limpieza y filtrado. A continuación se resumen los principales inconvenientes encontrados:

Cuadro 5: Principales problemas identificados en los datasets

Problema	**Descripción**
Valores atípicos extremos	Se identificaron distancias de viaje (`trip_distance`) mayores a 11 millones de millas y coordenadas geográficas fuera del rango real de la ciudad de Nueva York.
Tarifas y montos negativos	Existen registros con valores negativos en variables monetarias como `fare_amount`, `tip_amount` y `total_amount`, lo cual no es coherente con la lógica del sistema de transporte.
Ubicaciones geográficas erróneas	Varias trayectorias presentan coordenadas de recogida y destino ubicadas fuera del área metropolitana de NYC, afectando la validez del análisis espacial.
Valores nulos en variables clave	El dataset de sismos contiene valores nulos en columnas como `depth`, lo cual afecta el análisis de riesgo sísmico, especialmente en zonas urbanas donde la profundidad es un factor crítico.
Presencia de un registro duplicado	En el dataset de taxis se detectó un registro duplicado, que aunque estadísticamente insignificante, fue eliminado para mantener la coherencia estructural.

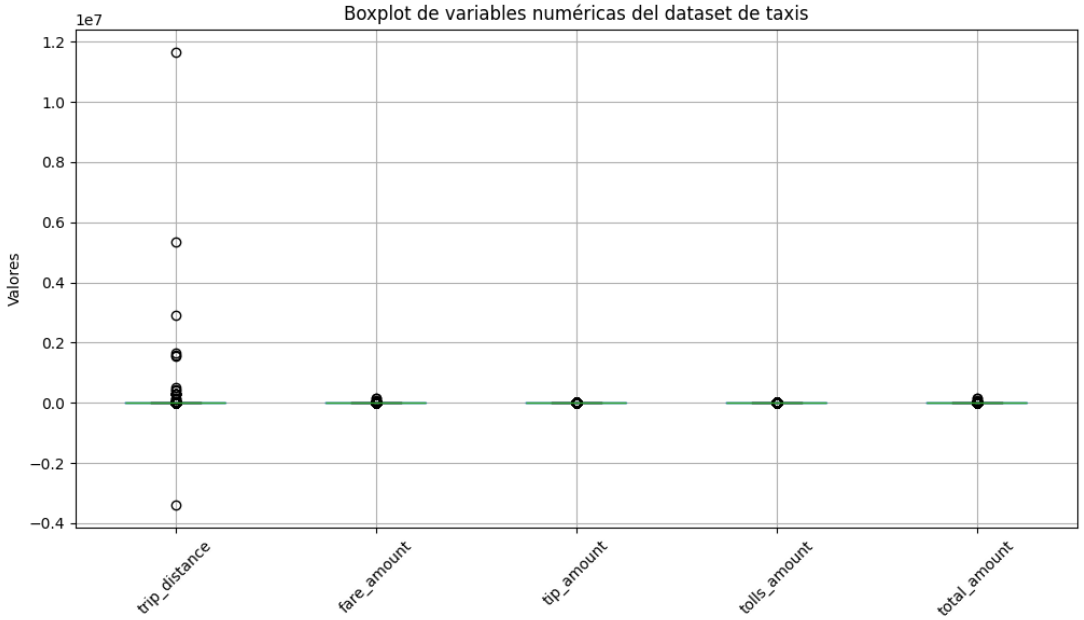

Figura 1: Detección de outliers en la variable `trip_distance` mediante gráfico BoxPlot.

Figura 2: Identificacion de las trayectorias fuera del rango Geografico real de Nueva York.

Estos problemas, si no se abordaban adecuadamente, podían distorsionar los resultados del análisis espacial y temporal, afectando la identificación correcta de zonas críticas y rutas vulnerables ante eventos sísmicos. Por ello, se aplicaron procesos rigurosos de depuración, cuyas acciones y efectos se detallan en las secciones de limpieza y transformación del presente estudio.

3. ¿Qué descubrieron al analizar los datos?

3.1. Alta concentración espacial de trayectorias en zonas céntricas

Se observa una clara concentración de puntos de recogida de taxi en zonas céntricas de la ciudad de Nueva York, especialmente en Manhattan. Esto refleja que dichas áreas tienen una alta demanda de transporte. Este patrón espacial es crucial para entender dónde se concentra la movilidad urbana y, por tanto, dónde un evento sísmico podría tener mayor impacto operativo y logístico.

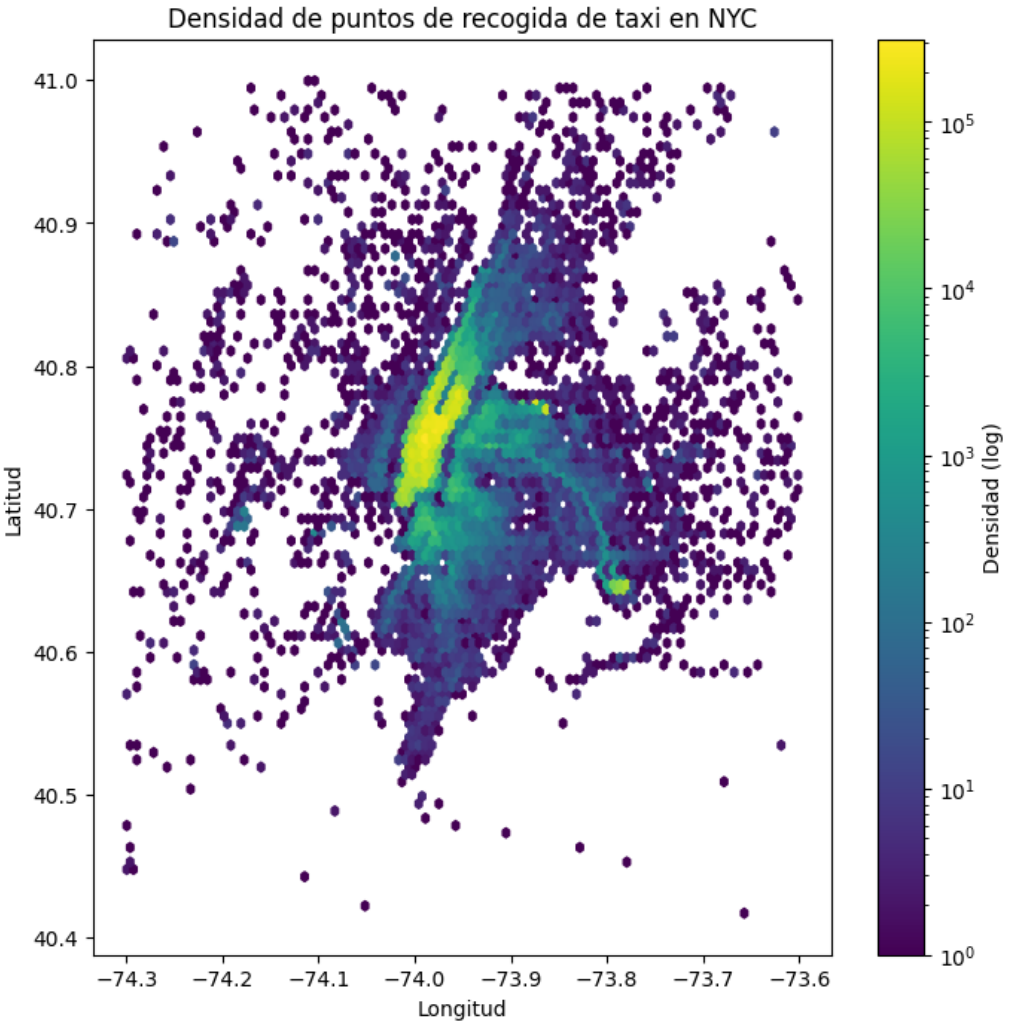

Figura 3: Mapa de densidad de puntos de recogida de los taxis

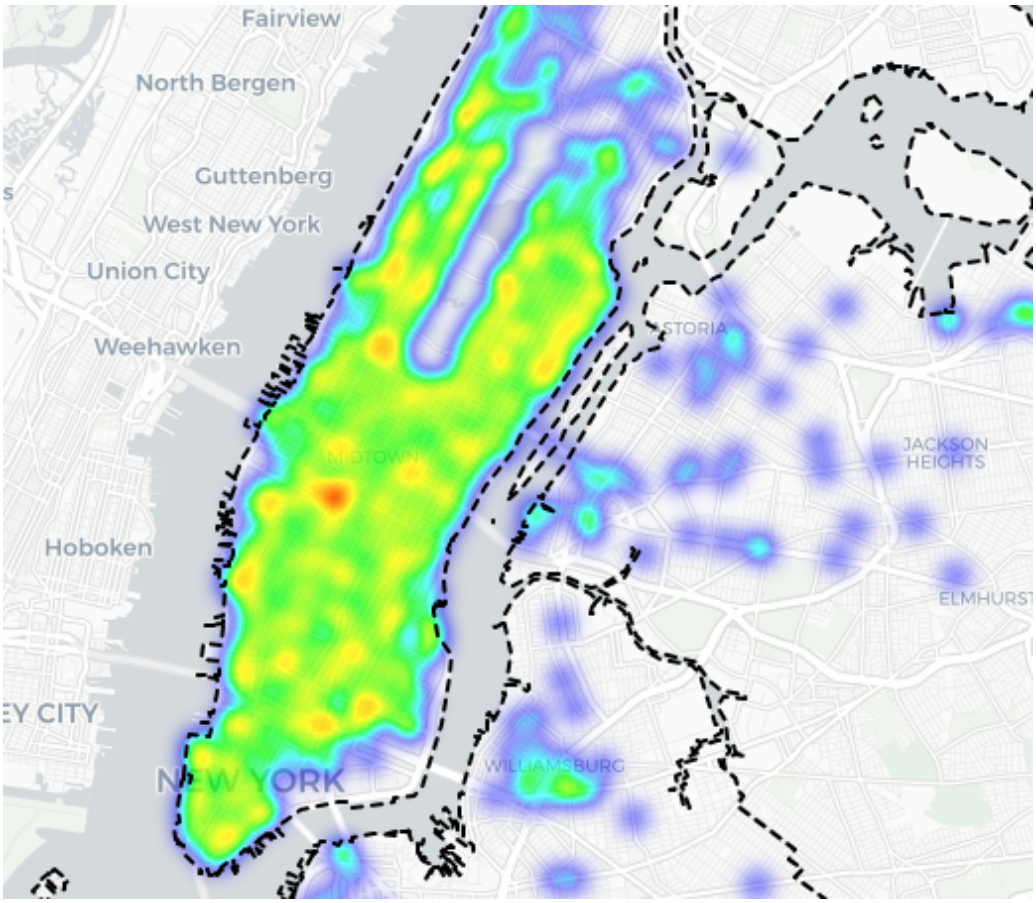

Figura 4: Mapa de densidad de puntos de recogida de taxi sobre los límites de la ciudad de Nueva York

Ambos descubrimientos permiten avanzar en la identificación de zonas vulnerables y evaluar los riesgos diferenciales según el espacio y el tiempo, contribuyendo a una planificación urbana resiliente frente a desastres naturales.

4. ¿Qué reflejan los patrones de tendencia?

El análisis de tendencias permitió identificar comportamientos temporales relevantes que fortalecen la comprensión del riesgo urbano frente a emergencias sísmicas, particularmente desde una perspectiva de movilidad. Se destacan dos patrones de interés:

1. Actividad urbana en ascenso frente a disminución de eventos sísmicos

Al comparar el número de recogidas de taxi con la cantidad de eventos sísmicos según franja horaria, se observa un comportamiento opuesto. La actividad urbana sigue una curva ascendente, con mayor número de viajes a medida que avanza el día, alcanzando su pico en la noche. En contraste, la frecuencia de sismos decrece progresivamente desde la mañana hasta la noche.

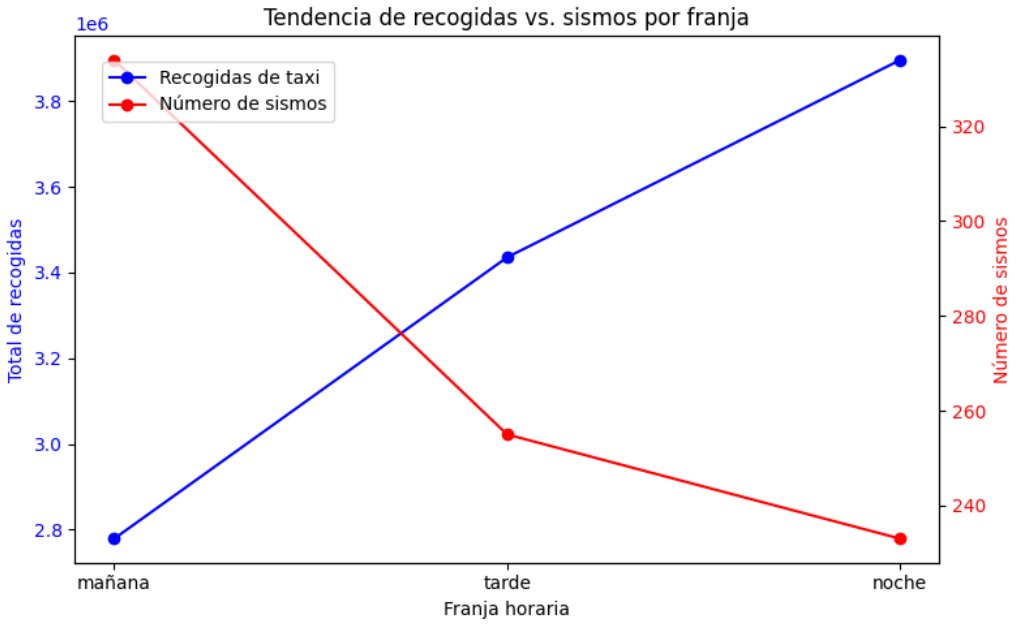

Figura 5: Tendencia comparativa de recogidas de taxi vs. eventos sísmicos por franja horaria

Este patrón sugiere que, aunque los sismos no son más frecuentes en horas de alta actividad urbana, cualquier evento ocurrido durante la noche (cuando hay más personas movilizándose) podría tener un mayor impacto operativo y logístico.

2. Trayectos más extensos en franjas de menor densidad

Un segundo hallazgo relevante es que la **distancia promedio de los trayectos** no se mantiene constante a lo largo del día. Se detecta una mayor distancia durante la **madrugada**, seguida de un descenso en la mañana, y luego una ligera recuperación en la tarde y noche.

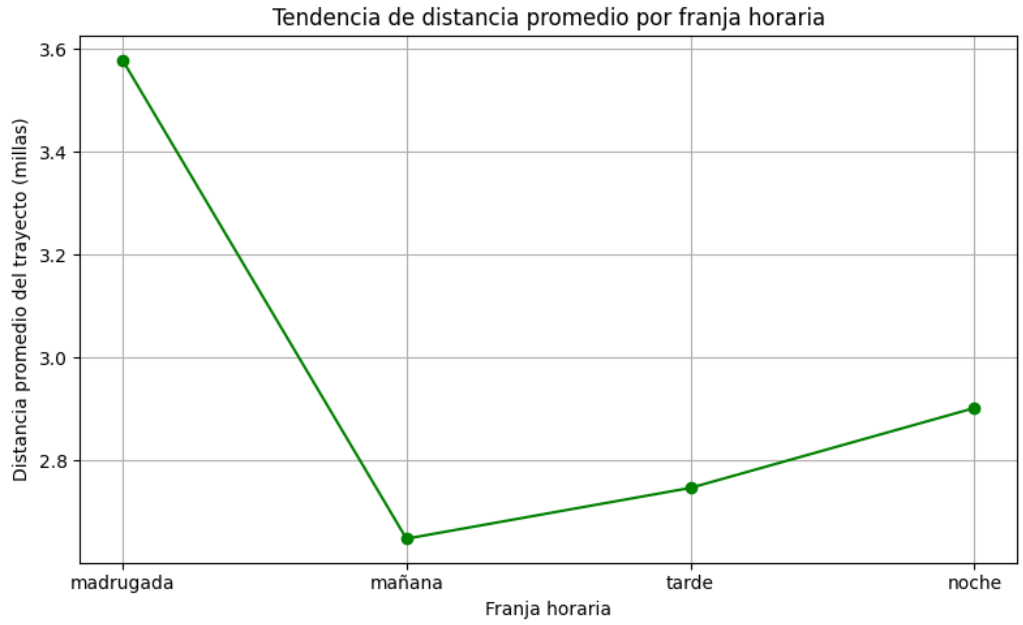

Figura 6: Tendencia de la distancia promedio de trayectos por franja horaria

Este comportamiento puede interpretarse como el predominio de trayectos largos hacia zonas periféricas o residenciales durante la madrugada (posiblemente tras jornadas laborales), mientras que en la mañana se concentran los trayectos cortos, típicos de desplazamientos laborales dentro de zonas céntricas.

Estos dos patrones refuerzan la necesidad de considerar tanto el tiempo como el tipo de trayectorias al planificar estrategias de evacuación, ya que las condiciones de movilidad varían significativamente a lo largo del día.

5. Pregunta planteada

¿Coinciden geográficamente las zonas de mayor actividad de movilidad urbana con las zonas donde han ocurrido más eventos sísmicos?

Esta pregunta busca integrar dos dimensiones clave del análisis: la actividad de movilidad urbana (trayectorias de taxi) y la exposición sísmica. Para responderla, se generó un mapa dinámico que superpone los puntos de recogida de taxi con los epicentros de eventos sísmicos registrados en la región de Nueva York.

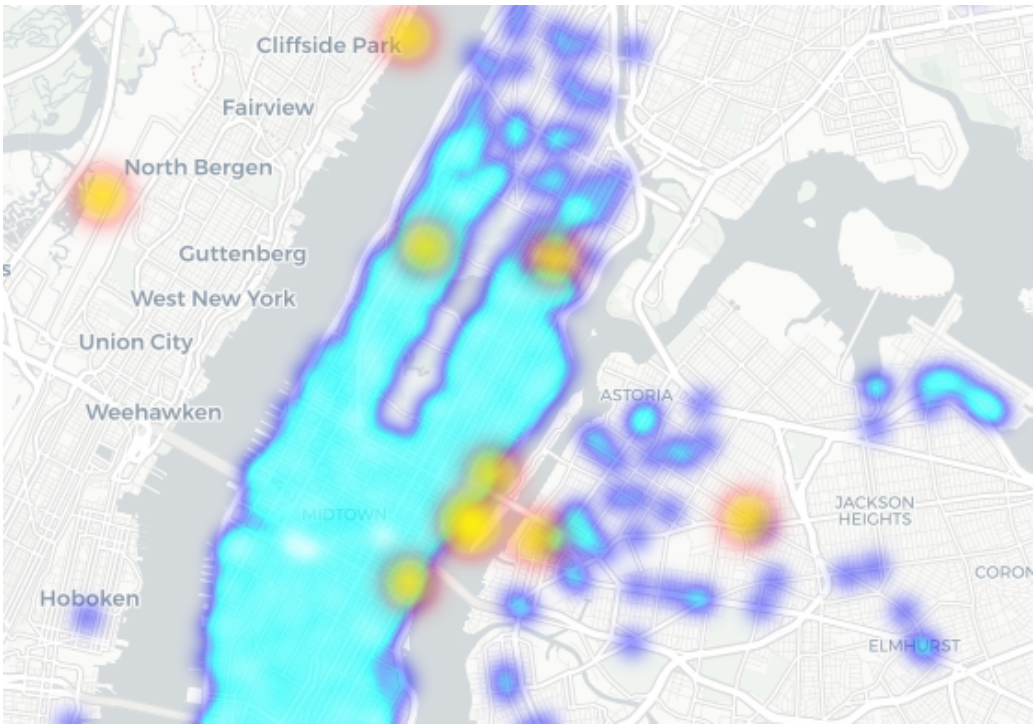

Figura 7: Superposición de densidad de trayectorias de taxi (azul) y eventos sísmicos (rojo-amarillo)

Interpretación

Como se aprecia en la figura, las zonas de mayor concentración de movilidad urbana —principalmente en **Manhattan, Midtown y áreas cercanas a Queens**— presentan también una significativa coincidencia con puntos de eventos sísmicos. Esta superposición espacial revela la existencia de **zonas críticas** donde no solo hay alta densidad de tránsito, sino también una mayor probabilidad histórica de actividad sísmica.

Estas áreas representan un doble riesgo en caso de desastre natural: la congestión por la alta demanda de transporte urbano, y la posible afectación directa por sismos. Este hallazgo refuerza la necesidad de diseñar **estrategias específicas de evacuación y respuesta rápida** en dichos sectores, integrando factores geoespaciales y patrones de movilidad.

Este tipo de análisis cruzado permite priorizar zonas vulnerables no solo por su exposición física al riesgo, sino también por su importancia funcional en la dinámica urbana diaria.