1. Título del Proyecto de Ciencia de Datos

Análisis y Visualización del Sistema de Transporte Público de San Francisco (SFMTA) a partir de Datos GTFS Estáticos

2. Contexto, Motivación y Justificación

El transporte público es muy importante para que las ciudades funcionen bien. Ayuda a que las personas se muevan de forma más ordenada, contamina menos que otros medios y reduce el tráfico. En los últimos años, muchas agencias de transporte han comenzado a publicar sus datos usando el formato GTFS (General Transit Feed Specification), que organiza información como rutas, horarios y paradas [Prommaharaj et al., 2020].

Aunque estos datos están disponibles, normalmente se usan solo para mostrar horarios o planificar recorridos básicos. Sin embargo, tienen mucho más potencial si se analizan con herramientas de visualización. Por ejemplo, G2Viz es una herramienta que usa datos GTFS para mostrar de forma visual cómo funciona el sistema de transporte, lo que ayuda a encontrar patrones y tomar mejores decisiones [Para et al., 2024].

Este proyecto busca usar los datos reales del transporte de San Francisco (SFMTA) para hacer un análisis exploratorio y crear visualizaciones que ayuden a entender cómo está organizado el sistema y cuál es su comportamiento.

3. Planteamiento del Problema

Aunque el formato GTFS es usado por muchas ciudades, está dividido en muchos archivos distintos y no es tan fácil de analizar si no se procesan bien primero. Esto puede ser un problema para estudiantes, ciudadanos o instituciones que quieran entender mejor el sistema de transporte [Wu et al., 2023].

Como señala [Para et al., 2024], cuando estos datos se visualizan de forma clara, es mucho más fácil notar cosas como cuáles zonas tienen más paradas, qué rutas son más frecuentes o en qué horarios hay mayor servicio. Aun así, no es común ver este tipo de análisis aplicado en profundidad a los datos de una ciudad.

Pregunta de investigación: *¿Cómo podemos transformar los datos GTFS estáticos de la SFMTA en información clara y visualmente comprensible sobre la estructura y operación del sistema de transporte público de San Francisco?*

4. Objetivo General

Analizar y visualizar el sistema de transporte público de San Francisco (SFMTA) a partir de los datos GTFS estáticos, para entender mejor cómo está estructurado y cómo funciona el servicio.

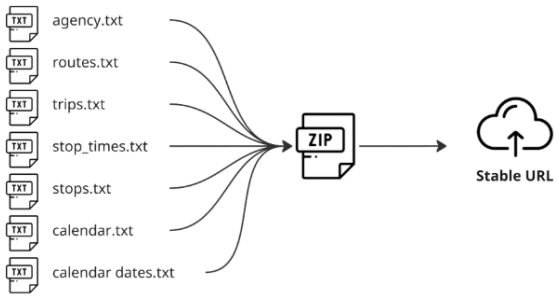
5. Objetivos Específicos

- Explorar la estructura de los archivos que componen el dataset GTFS de SFMTA.
- Limpiar y preparar los datos para que puedan ser analizados y visualizados.
- Identificar la cantidad de rutas, los tipos de transporte y cuándo están activos.
- Analizar la frecuencia de los servicios y la cobertura geográfica del sistema.
- Crear visualizaciones que ayuden a interpretar los patrones de uso del sistema de transporte.

6. Descripción del Dataset

Dataset:muni_gtfs-current.zip

El dataset utilizado contiene datos GTFS actualizados de la **San Francisco Municipal Transportation Agency (SFMTA)**. Este conjunto incluye archivos como `routes.txt`, `stops.txt`, `trips.txt`, `stop_times.txt`, entre otros. Cada archivo representa una entidad o relación específica dentro del sistema de transporte, y su estructura permite consultas cruzadas para responder preguntas sobre operación y cobertura del servicio.



Cada archivo representa una entidad clave del sistema (por ejemplo, una parada o una ruta) o una relación entre entidades (como la secuencia de paradas de un viaje). Esta organización modular facilita el cruce de información y permite responder preguntas complejas sobre el funcionamiento operativo del sistema.

El dataset contiene 12 archivos principales que difieren en tamaño, codificación y nivel de granularidad. Archivos como `stop_times.txt` y `trips.txt` presentan gran volumen de registros y detalles temporales finos, mientras que archivos auxiliares como `calendar.txt` o `fare_rules.txt` son más compactos. La mayoría están codificados en ASCII o UTF-8, lo que facilita su lectura mediante herramientas estándar como Pandas.

Archivo	Codificación	Tamaño (MB)	Granularidad
agency.txt	ascii	0.000000	1 filas x 8 columnas
stops.txt	ascii	0.272000	3278 filas x 8 columnas
routes.txt	ascii	0.006000	69 filas x 10 columnas
trips.txt	ascii	1.885000	35017 filas x 9 columnas
stop_times.txt	ascii	54.836000	1313881 filas x 9 columnas
calendar.txt	ascii	0.000000	3 filas x 10 columnas
calendar_dates.txt	ascii	0.000000	26 filas x 3 columnas
fare_attributes.txt	ascii	0.000000	2 filas x 7 columnas
fare_rules.txt	ascii	0.000000	69 filas x 2 columnas
shapes.txt	ascii	1.798000	46249 filas x 5 columnas
directions.txt	ascii	0.002000	138 filas x 3 columnas
timepoints.txt	ascii	4.527000	249904 filas x 2 columnas

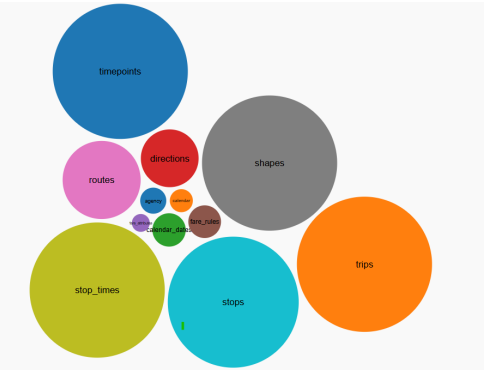


Figura 1: Visualización de archivos GTFS: tamaño y granularidad.

Modelo relacional Para entender cómo se interconectan los datos, se elaboró un **modelo entidad-relación (ER)** que representa las claves primarias y foráneas usadas para vincular entidades. Por ejemplo, `trip_id`, `stop_id`, `route_id` y `shape_id` permiten reconstruir trayectos, calcular frecuencias o analizar rutas espacialmente.

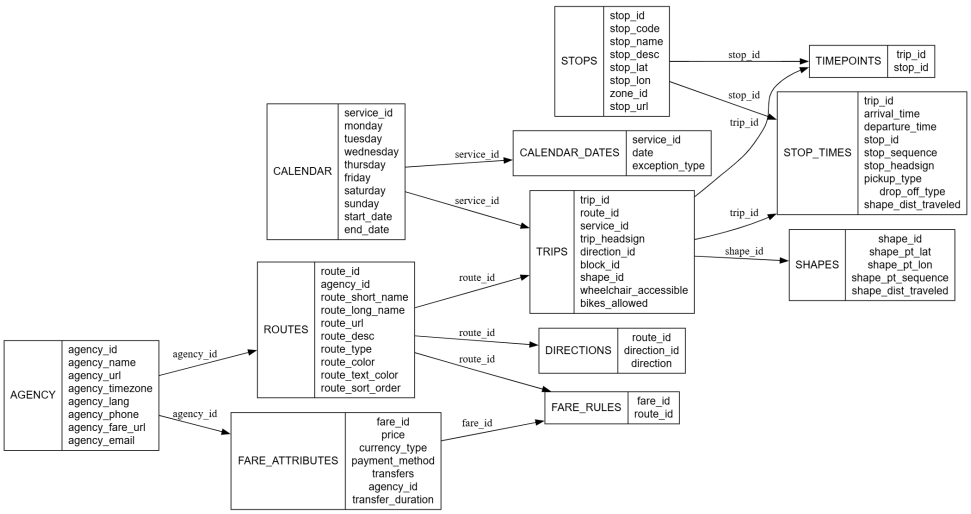


Figura 2: Modelo relacional de archivos incluidos en el GTFS SFMTA.

Exploración por archivo Durante la etapa exploratoria, cada archivo fue examinado individualmente:

- `agency.txt`: define la agencia operadora (SFMTA). Campos como `agency_timezone` y `agency_lang` son críticos para interpretar correctamente los horarios y los textos en apps.
- `stops.txt`: lista las paradas con coordenadas geográficas. Un mismo nombre puede aparecer con distintos `stop_id` si representa diferentes sentidos de viaje.
- `routes.txt`: describe las rutas activas, su tipo (bus, tranvía, cable car), colores y variantes como rutas expresas o nocturnas.

- `trips.txt`: contiene cada viaje programado, incluyendo dirección (`direction_id`), accesibilidad y `shape_id` del recorrido.
- `stop_times.txt`: vincula viajes con paradas y sus horarios. Soporta tiempos extendidos como `25:00:00` para representar servicios nocturnos.
- `calendar.txt` y `calendar_dates.txt`: definen la operación regular y las excepciones (como feriados o eventos especiales).
- `fare_rules.txt` y `fare_attributes.txt`: describen las reglas tarifarias aplicadas por ruta. SFMTA utiliza una tarifa plana simple.
- `shapes.txt`: registra la geometría de las rutas para su representación cartográfica.
- `directions.txt`: traduce los valores 0/1 en etiquetas comprensibles como "Inbound" y "Outbound".
- `timepoints.txt`: lista paradas clave utilizadas como puntos de control horario.

Análisis de calidad de datos Se detectaron campos con valores nulos que, si bien no afectan la integridad estructural, podrían limitar visualizaciones o simulaciones:

Archivo	**Columnas afectadas**	**Descripción del problema**
`stops.txt`	`stop_desc`, `zone_id`	Muchas filas sin descripción o sin zona tarifaria asignada.
`stop_times.txt`	`stop_headsign`	Este campo suele estar vacío si no hay información adicional.
`routes.txt`	`route_sort_order`	Campo vacío que podría usarse para ordenar rutas en visualizaciones.
`fare_attributes.txt`	`transfers`, `agency_id`	Faltan valores clave para modelar transferencias o no aplican al caso de una sola agencia.

Cuadro 1: Resumen de columnas con valores nulos detectadas en archivos GTFS

Limpieza y transformación Inspirado en el enfoque de **G2Viz**, se ejecutó un proceso de limpieza para optimizar el dataset para análisis visual. Esto incluyó:

- Eliminación de columnas irrelevantes (como `stop_url`, `block_id`, etc.).
- Creación de versiones depuradas:
  - `stops.csv`: paradas con solo campos esenciales.
  - `trips.csv`: estructura reducida útil para análisis de frecuencia y trayectorias.

	Tabla	Filas antes	Columnas antes	Filas después	Columnas después	Columnas eliminadas
0	stops	3278	8	3278	6	zone_id, stop_url
1	trips	35017	9	35017	5	wheelchair_accessible, bikes_allowed, block_id, direction_id

Figura 3: Comparación de tablas originales vs limpias.

Este preprocesamiento facilita el análisis exploratorio, mejora la eficiencia computacional y permite desarrollar dashboards interactivos o simulaciones espaciales y temporales del sistema de transporte.

Contexto del Dataset

El formato **GTFS (General Transit Feed Specification)** fue desarrollado originalmente por Google y TriMet (la agencia de transporte de Portland) con el objetivo de estandarizar la publicación de datos de transporte público. Este estándar abierto es utilizado por agencias de todo el mundo para compartir información detallada sobre horarios, rutas, paradas y servicios de transporte.

Gracias a su estructura tabular y relacional, los datos GTFS pueden ser procesados mediante herramientas de análisis de datos y sistemas de información geográfica (SIG), permitiendo realizar estudios sobre cobertura del servicio, eficiencia operativa y planificación del transporte.

El archivo `muni_gtfs-current.zip` representa una versión actualizada del sistema de transporte público de San Francisco. Contiene los servicios programados al momento de la descarga, incluyendo todos los viajes diarios planificados y sus respectivas características espaciales y temporales.

7. Preguntas

A partir de los objetivos específicos definidos en este estudio y el análisis visual realizado sobre el dataset GTFS de SFMTA, se formularon las siguientes preguntas clave. Cada una de ellas está orientada a responder aspectos operativos, espaciales o estructurales del sistema de transporte, y fue explorada mediante visualizaciones específicas construidas a partir de los archivos GTFS procesados:

Cuadro 2: Relación entre preguntas, objetivos específicos, archivos GTFS y propósito analítico

Pregunta	**Objetivo Específico**	**Archivo(s) GTFS**	**¿Qué se busca analizar o descubrir?**
¿Cuántas rutas existen y de qué tipo son (bus, tranvía, cable car)?	Identificar la cantidad de rutas, los tipos de transporte y cuándo están activos.	`routes.txt`	Conocer la variedad de medios operados por SFMTA y su distribución en el sistema.
¿Cuáles son las rutas con más viajes programados por día?	Analizar la frecuencia de los servicios.	`trips.txt`	Detectar rutas de alta frecuencia. Refleja demanda o prioridad operativa.
¿Qué zonas tienen más cobertura o mayor densidad de paradas?	Analizar la cobertura geográfica del sistema.	`stops.txt`	Evaluar la accesibilidad del sistema por zonas y distribución espacial de paradas.
¿Las rutas más largas (con más paradas) son también las más frecuentes?	Analizar la frecuencia de los servicios y su estructura.	`stop_times.txt`, `trips.txt`	Contrastar longitud y frecuencia de rutas para entender decisiones operativas.
¿Qué días y franjas horarias tienen más servicio?	Crear visualizaciones para interpretar patrones de uso del sistema de transporte.	`calendar.txt`, `trips.txt`, `stop_times.txt`	Descubrir diferencias entre días de semana y fines de semana, y patrones horarios como las horas pico.

Estas preguntas estructuran el análisis exploratorio y permiten alinear cada hallazgo visual con una interrogante concreta del sistema de transporte público de San Francisco.

Referencias

[Para et al., 2024] Para, S., Wirotsasithon, T., Jundee, T., Demissie, M. G., Sekimoto, Y., Biljecki, F., and Phithakkitnukoon, S. (2024). G2viz: An online tool for visualizing and analyzing a public transit system from gtfs data. *Public Transport*, 16(3):893–928.

[Prommaharaj et al., 2020] Prommaharaj, S., Kattan, L., Shah, S., and Farooq, B. (2020). Visualizing public transit system operation with gtfs data: A case study of calgary, canada. *Heliyon*, 6(5):e04036.

[Wu et al., 2023] Wu, J., Du, B., Gong, Z., Wu, Q., Shen, J., Zhou, L., and Cai, C. (2023). A gtfs data acquisition and processing framework and its application to train delay prediction. *International Journal of Transportation Science and Technology*, 12(1):201–216.