Milestone 01

> Data Abstraction for Citi Bike Trip Histories

Group: 11
GitHub Repository: https://github.com/tsaichen1o/jc-citibike-vis

A. Title & Source

- **Dataset Title:** Citi Bike NYC System Data (JC-202509-citibike-tripdata)
- **Primary URL:** https://citibikenyc.com/system-data
- **Publisher/Author:** Citi Bike NYC
- **Publication/Last Update Date:** Oct 6th 2025
- **License/Usage Terms:** NYCBS Data Use Policy

B. Motivation

The Citi Bike dataset interests us because it captures the dynamic rhythms of urban mobility and reveals how people navigate shared infrastructure. Its rich temporal and spatial dimension make it ideal for exploring behavioral patterns and optimizing transportation systems. Two questions we aim to explore are: 1. How do member and casual users differ in their riding activity across hours of the day? 2. What are the most common origin-destination routes based on geographic coordinates?

C. Scope & Granularity

Our dataset contains about 100k records, each with 13 attributes. Every record represents a single bike trip. The `ride_id` serves as the unique identifier for each trip, while `start_station_id` and `end_station_id` are key fields for linking the start and end points of the journeys.

D. Schema (Types & Ranges)

Attribute	Role	Type	T/S	Domain/Range
`ride_id`	ID/Key	Cat.	N	~110k unique IDs
`rideable_type`	Attr.	Cat.	N	{'classic_bike', ...}
`started_at`	Attr.	Quant.	Y (T)	Datetime
`ended_at`	Attr.	Quant.	Y (T)	Datetime
`start_station_name`	Attr.	Cat.	N	~100 stations
`start_station_id`	ID/Key	Cat.	N	e.g., 'JC005'

Attribute	Role	Type	T/S	Domain/Range
`end_station_name`	Attr.	Cat.	N	~100 stations
`end_station_id`	ID/Key	Cat.	N	e.g., 'JC008'
`start_lat`	Attr.	Quant.	Y (S)	40.69-40.75
`start_lng`	Attr.	Quant.	Y (S)	-74.10 to -74.02
`end_lat`	Attr.	Quant.	Y (S)	(Similar)
`end_lng`	Attr.	Quant.	Y (S)	(Similar)
`member_casual`	Attr.	Cat.	N	{'member', 'casual'}

F. Quality & Limitations

The main data quality issues are missing values and outliers. Some trips, particularly those with electric bikes, have no end station information because they can be parked anywhere; this is a real-world scenario we need to handle in our visualization. We also expect to find outliers in trip duration. To prepare the data, we will first calculate trip duration from the start and end times. Then, we will filter out very short trips (under 60 seconds) and very long ones (over 24 hours) to remove potential errors, and we will explicitly manage trips with missing end-station data.

G. Suitability

This dataset comes as a single CSV file, which makes it easy to start working with without much cleanup. However, with 100k records, it is large enough to reveal interesting patterns in urban mobility and to test the performance of our D3.js visualizations. The data is also rich, combining temporal, spatial, and network information. This variety allows us to create multiple types of visualizations, such as flow maps to explore and compare trip data effectively.