Data Sources Report

Data Analyst: Joshua Valdez

Client/Sponsor: Cyclistics - Marketing Strategy Department

Date: July 9th, 2025

Goal: Clearly and transparently describe all data sources used in the analysis, their format, origin, prior processing, and inclusion criteria.

1. Data overview

The data used in this project corresponds to historical records of bicycle trips publicly provided by the Divvy bike-sharing system in Chicago, which is referred to as Cyclistics in this fictional case study.

- Data type: .csv files (comma-separated values)
- Frequency: Monthly
- Period covered: January 2024 – December 2024
- Total files: 12
- Approximate total size: ~1.01 GB
- Language: English
- License: Open Data Commons Public Domain Dedication and License (PDDL)

2. Download source

- All files were downloaded from the official Divvy website at the following URL:

  https://divvy-tripdata.s3.amazonaws.com/index.html

- Each file follows the format:

  YYYYMM-divvy-tripdata.csv

- Example:
  - 202401-divvy-tripdata.csv
  - 202402-divvy-tripdata.csv
  - ...
  - 202412-divvy-tripdata.csv

3. Fields included in the files

Standardized fields after cleaning include:

Original Column	Description
ride_id	Unique trip identifier
rideable_type	Type of bike used (classic, electric, electric scooter)
started_at	Start date and time of the trip
ended_at	End date and time of the trip
start_station_name	Start station name
end_station_name	Destination station name
start_station_id	Start station unique identification code
end_station_id	Destination station unique identification code
start_lat	Start latitude
start_lng	Start longitude
end_lat	Destination latitude
end_lng	Destination longitude
member_casual	User type: member or casual

During the prepare phase, the following additional variables were created:

New Variable	Description
ride_length_sec	Ride length in seconds
ride_length_min	Ride length in minutes
day_of_week	Day of the week
month	Month in which the trip occurred
date	Clean date for daily calculations

4. Data integrity validation

Before any analysis were done, all files were reviewed to:

- Ensure column and format matching
- Eliminate records with:
  - Negative or zero trip duration
  - Incorrectly formatted dates
  - Critical empty values (ride_id, started_at, ended_at, member_casual)
- Standardize data types to avoid join conflicts (bind_rows())

5. Considerations and exclusions

- The March, April, May, and June 2024 files had a high percentage of missing values in station fields, but this was retained because the time and user data were complete.
- Unused variables such as ride_id, start_station_id, and end_station_id were excluded from the analysis, as they are no longer relevant under the new Cyclistics structure.
- The files were then saved into .xlsx files for better manipulation in R.

6. Final observations

The data used complies with ROCCC principles:

- Reliable: official and maintained source
- Original: generated by the official operator
- Comprehensive: covers all trips during the period
- Current: complete data for 2024
- Cited: license and download source correctly referenced