Data Cleaning Report

Data Analyst: Joshua Valdez

Client/Sponsor: Cyclistics - Marketing Strategy Department

Date: July 9th, 2025

Goal: Describe in detail the data cleaning process used in the analysis, including technical steps and criteria applied to ensure the integrity of the final dataset

1. File unification

Twelve CSV files were collected corresponding to each month of the year 2024. Each file was inspected to verify that it contained the following key columns:

- ride_id
- rideable_type
- started_at
- ended_at
- start_station_name
- end_station_name
- member_casual

The files were loaded and combined into a single dataframe using R, ensuring consistency in column names through the `clean_names()` function in the `janitor` package.

2. Data type conversion

The date and time columns (`started_at`, `ended_at`) have been converted to POSIXct type using the `ymd_hms()` function from the `lubridate` package to allow precise time operations.

The following derived variables have also been added:

- ride_length_sec: Difference between ended_at and started_at in seconds.
- ride_length_min: Difference in minutes (ride_length_sec / 60).
- day_of_week: Day of the week the ride occurred.
- month: Month in which the ride occurred.

- date: Simplified date (without time) for daily groupings.

3. Removing problematics columns

Some files contained additional columns such as `ride_length` with inconsistent formats (text, time). These columns were removed before combining the files to avoid type errors in the `bind_rows()` function.

4. Filtering invalid records

The following filters were applied to ensure integrity and quality:

- Rows with NA values for ride_id, started_at, ended_at, or member_casual were eliminated.
- Trips with negative or zero durations were eliminated.
- Rows with invalid date formats were excluded.

Records without stations (`start_station_name` or `end_station_name` being empty) were retained if the time and user data were complete.

5. Final validation

Data types were checked with `glimpse()` and parsing issues were verified using `problems()` from the `readr` package. Implicit duplicates were also removed, and the `rideable_type` column was ensured to be present and correctly categorized. The final dataset structure was confirmed and saved for use in subsequent exploratory analyses in R and SQL.