# Capital Bikeshare Cleaned Data README

## Scope

This README documents the cleaned datasets produced by the notebook clean_code.ipynb. It does not describe raw inputs, scraping infrastructure, or other project artifacts beyond the cleaned outputs.

## Coverage

- Geography: Capital Bikeshare service area (DC, Arlington, Alexandria, Montgomery County, Fairfax, Prince George's County, Falls Church)
- Main time window for station-level and free-bike feeds: 2019-04-01 through 2022-09-30
- Trip data window: starts from 2019-04-01; ended_at values extend into early October 2022 for trips that began near the end of September

## Cleaned Data Products

All cleaned products are stored in Data/Clean.

### File Naming Convention

For Station Information, Station Status, Free Bike Status, and Trips, each subfolder contains multiple files, with one file per month.

- CSV subfolders contain monthly `.csv` files.
- Parquet subfolders contain monthly `.parquet` files.
- File names follow a `YYYY-MM-<dataset_name>.<ext>` pattern.

Example: - `Station Information - csv` contains monthly files named from `2019-04-station_information.csv` through `2022-09-station_information.csv`. - `Station Information - parquet` uses the same monthly naming pattern with `.parquet` extension.

The same month-by-month structure is used for Station Status, Free Bike Status, and Trips, with dataset-specific suffixes.

1. Station Information
2. Location: Data/Clean/Station Information
3. Files: monthly CSV files and monthly Parquet files
4. File coverage: monthly files (`2019-04` through `2022-09`), one file per month in each subfolder
5. Unit of observation: station by scrape minute
6. Primary identifiers: station_id, short_name, scrape_time

7. Description: cleaned station metadata over time (name, coordinates, capacity, region, kiosk/key dispenser indicators), with station ID and naming inconsistencies harmonized
8. Station Status
9. Location: Data/Clean/Station Status
10. Files: monthly CSV files and monthly Parquet files
11. File coverage: monthly files (`2019-04` through `2022-09`), one file per month in each subfolder
12. Unit of observation: station by scrape minute
13. Primary identifiers: station_id, scrape_time
14. Description: cleaned station availability/status counts (bikes, docks, e-bikes, disabled counts, operational flags), aligned to cleaned station identifiers
15. Free Bike Status
16. Location: Data/Clean/Free Bike Status
17. Files: monthly CSV files and monthly Parquet files
18. File coverage: monthly files, one file per month in each subfolder (coverage begins in `2020-06`)
19. Unit of observation: bike_id by scrape minute
20. Primary identifiers: bike_id, scrape_time
21. Description: cleaned free-floating bike records (location and reservation status)
22. Note: free-bike observations begin in mid-2020 in this dataset
23. Trips
24. Location: Data/Clean/Trips
25. Files: monthly CSV files and monthly Parquet files
26. File coverage: monthly files (`2019-04` through `2022-09`), one file per month in each subfolder
27. Unit of observation: trip
28. Description: cleaned monthly trip records with harmonized schema across historical format changes
29. System Regions
30. Location: Data/Clean
31. Files: system_regions.csv and system_regions.parquet
32. Unit of observation: region_id
33. Primary identifier: region_id
34. Description: cleaned lookup table from region_id to region name, restricted to regions used in cleaned station information
35. Gap Files
36. Location: Data/Clean
37. Files: Station Information Gaps.csv and Station Status Gaps.csv
38. Unit of observation: contiguous scrape-time gap window
39. Description: summaries of missing scrape intervals in the station-level feeds, including the interval start/end and number of stations impacted
40. DC Ward Boundaries (GeoJSON)
41. Location: Data/Clean
42. Files: Wards_from_2022.geojson
43. Unit of observation: ward polygon feature

44. Description: GeoJSON shapefile defining Washington, DC ward boundaries (from 2022 onward) for mapping and spatial joins

# Key Variables by Dataset

Below are the key analytic fields expected after cleaning. Some optional fields appear only in subsets of months, consistent with source feed changes.

1. Station Information
2. station_id, short_name, name
3. lat, lon
4. region_id
5. capacity
6. has_kiosk, has_key_dispenser
7. last_updated, ttl, scrape_time
8. same_as_previous, same_as_previous_excl_last_updated
9. optional station_services_* fields in periods where valet data exists
10. Station Status
11. station_id
12. num_docks_available, num_bikes_available, num_ebikes_available
13. num_docks_disabled, num_bikes_disabled
14. is_installed, is_returning, is_renting
15. has_available_keys
16. last_reported, last_updated, ttl, scrape_time
17. same_as_previous, same_as_previous_excl_last_updated
18. optional valet_* fields in periods where valet data exists
19. Free Bike Status
20. bike_id
21. lat, lon
22. is_reserved
23. last_updated, ttl, scrape_time
24. same_as_previous, same_as_previous_excl_last_updated
25. Trips
26. started_at, ended_at
27. start_station_id, end_station_id (these correspond to station short_name values)
28. member_casual (coded to numeric)
29. rideable_type (numeric coding where present)
30. ride_id and bike_number may appear depending on source month/schema
31. System Regions
32. region_id
33. name
34. Gap Files
35. start_gap
36. stop_gap
37. num_stations_impacted
38. stations_impacted

39. DC Ward Boundaries (GeoJSON)
40. geometry (polygon/multipolygon)
41. ward identifier and ward name attributes (as provided in the source GeoJSON)

# Join Guidance

1. Station status to station information
2. Recommended join keys: station_id plus aligned time logic using scrape_time
3. Trips to station information
4. start_station_id and end_station_id map to short_name in Station Information (not station_id)
5. Station information to system regions
6. Join on region_id

# Cleaning Outcomes Included in These Files

The cleaned outputs include: - removal of known invalid/private/test records - harmonization of known station identifier changes - correction of many known station name typos/inconsistencies - targeted imputation/fixes for missing or zeroed coordinates and missing region values where recoverable - type standardization for dates, numeric counts, and boolean flags - duplicate flag construction using same_as_previous and same_as_previous_excl_last_updated - filtering of invalid trip durations and exact duplicate trips

# Important Caveats

- Some fields are structurally missing in early months because upstream APIs/files changed over time.
- A subset of station records still has capacity equal to zero; these were retained.
- Free bike IDs are privacy-preserving and rotate frequently, so they should not be treated as persistent bike identifiers over long horizons.