

Package ‘sfOpenData’

September 8, 2026

Type Package

Title A Lightweight Interface to San Francisco Open Data APIs

Version 0.1.2

Description Provides a unified set of helper functions to access datasets from the San Francisco Open Data platform <<https://data.sfgov.org/>>. Functions return results as tidy tibbles and support optional filtering, sorting, and row limits via the Socrata API. The package provides a consistent interface for discovering and downloading datasets from the San Francisco Open Data Portal using human-readable dataset keys or official Socrata dataset identifiers.

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 8.0.0

Imports curl, dplyr, httr, janitor, jsonlite, rlang, tibble

Suggests ggplot2, knitr, rmarkdown, scales, testthat (>= 3.0.0), tidyrr, vcr (>= 0.6.0), webmockr

URL <https://github.com/nyc-open-data-lab/sfOpenData>,
<https://github.com/nyc-open-data-lab/sfOpenData/>

BugReports <https://github.com/nyc-open-data-lab/sfOpenData/issues>

VignetteBuilder knitr

Config/testthat/edition 3

Depends R (>= 4.1.0)

NeedsCompilation no

Author Shelby Lyn Gomes [aut] (GitHub: gomes-sh),
Christian Martinez [aut, cre] (GitHub: martinezc1, ORCID:
<<https://orcid.org/0009-0005-6026-6454>>)

Maintainer Christian Martinez <c.martinez0@outlook.com>

Repository CRAN

Date/Publication 2026-09-08 12:30:11 UTC

Contents

sf_any_dataset	2
sf_list_datasets	3
sf_pull_dataset	4

Index	8
--------------	----------

sf_any_dataset	<i>Pull any San Francisco Open Data dataset from a Socrata JSON endpoint</i>
----------------	--

Description

Downloads data from any San Francisco Open Data Socrata JSON endpoint and returns the result as a tibble. This function is useful for datasets that are not included in the curated catalog returned by [sf_list_datasets()].

Usage

```
sf_any_dataset(
  json_link,
  limit = 10000,
  timeout_sec = 30,
  clean_names = TRUE,
  coerce_types = TRUE
)
```

Arguments

json_link	A single Socrata dataset JSON endpoint URL, such as <code>"https://data.sfgov.org/resource/abcd-1234.json"</code> .
limit	Number of rows to retrieve. Defaults to 10,000.
timeout_sec	Request timeout in seconds. Defaults to 30.
clean_names	Logical. If <code>'TRUE'</code> , column names are converted to snake_case using [janitor::clean_names()]. Defaults to <code>'TRUE'</code> .
coerce_types	Logical. If <code>'TRUE'</code> , the package attempts lightweight, heuristic-based type coercion after downloading the data. Columns are converted only when at least 95 percent of non-missing values can be parsed as the target type. This helps avoid unsafe conversions when source data are inconsistent.

Details

San Francisco Open Data datasets have Socrata JSON endpoints that usually follow this pattern:

`'https://data.sfgov.org/resource/<dataset_ID>.json'`

For example, a dataset with the Socrata ID `"wr8u-xric"` would have the JSON endpoint:

```
'https://data.sfgov.org/resource/wr8u-xric.json'
```

Users can find a dataset's ID from the San Francisco Open Data Portal URL, the API documentation page for the dataset, or the output of `[sf_list_datasets()]`.

`'sf_any_dataset()'` bypasses the package catalog and sends a request directly to the supplied JSON endpoint. Unlike `[sf_pull_dataset()]`, it does not look up defaults such as catalog keys, date fields, or default ordering.

This function is intended for direct endpoint access. For catalog-based workflows using readable keys or Socrata IDs, use `[sf_pull_dataset()]`.

Value

A tibble containing rows from the requested San Francisco Open Data endpoint.

Examples

```
# Examples that hit the live San Francisco Open Data API are guarded so CRAN
# checks do not fail when the network is unavailable or slow.
if (interactive() && curl::has_internet()) {
  # Build a JSON endpoint from a Socrata ID
  ID <- "abcd-1234"
  endpoint <- paste0("https://data.sfgov.org/resource/", ID, ".json")

  out <- try(sf_any_dataset(endpoint, limit = 3), silent = TRUE)
  if (!inherits(out, "try-error")) {
    head(out)
  }
}
```

`sf_list_datasets`

List available San Francisco Open Data datasets

Description

Retrieves a live catalog of San Francisco Open Data datasets from the San Francisco Open Data Portal catalog table (`'CATALOG_DATASET_ID'`) and returns dataset metadata that can be used with `[sf_pull_dataset()]`.

Usage

```
sf_list_datasets()
```

Details

The returned tibble includes both the official Socrata dataset ‘ID’ and a human-readable ‘key’. The ‘ID’ is the stable identifier used by the San Francisco Open Data Portal and Socrata API, while the ‘key’ is generated from the dataset name using `[janitor::make_clean_names()]` to make datasets easier to reference in R code, especially in classroom and reproducible research settings.

Most users will begin by calling ‘sf_list_datasets()’, searching the returned catalog for a dataset of interest, and then passing either the ‘key’ or ‘ID’ to `[sf_pull_dataset()]`.

Value

A tibble containing available San Francisco Open Data datasets. Important columns include:

key A human-readable dataset key generated from the dataset name.

dataset_id The official Socrata dataset identifier used by San Francisco Open Data.

dataset_name The dataset name from the San Francisco Open Data catalog.

Additional metadata columns from the San Francisco Open Data catalog may also be returned.

Examples

```
if (interactive() && curl::has_internet()) {
  catalog <- sf_list_datasets()

  # View available datasets
  catalog

  # Search for datasets containing a keyword
  catalog[
    grepl("Health", catalog$dataset_name, ignore.case = TRUE),
    c("key", "dataset_id", "dataset_name")
  ]
}
```

sf_pull_dataset

Pull a San Francisco Open Data dataset

Description

Downloads a dataset from the San Francisco Open Data Socrata API using either a human-readable catalog ‘key’ or the official Socrata dataset ‘dataset_id’ returned by `[sf_list_datasets()]`.

Usage

```
sf_pull_dataset(
  dataset,
  limit = 10000,
  filters = list(),
  date = NULL,
  from = NULL,
  to = NULL,
  date_field = NULL,
  where = NULL,
  order = NULL,
  timeout_sec = 30,
  clean_names = TRUE,
  coerce_types = TRUE
)
```

Arguments

dataset	A single dataset 'key' or Socrata dataset 'dataset_id' from [sf_list_datasets()]. For example, a key may look like "example_dataset_name", while a dataset may look like "abcd-1234".
limit	Number of rows to retrieve. Defaults to 10,000.
filters	Optional named list of exact-match filters. Each list name should be a field name in the dataset, and each value should be the value or values to match. Vector values are translated into SQL-style 'IN' conditions in the generated SoQL query. For example, 'filters = list(FILTER_FIELD = c("VALUE_1", "VALUE_2"))' returns rows where 'FILTER_FIELD' is either "VALUE_1" or "VALUE_2".
date	Optional single date used to match all records from that day. Requires 'date_field'.
from	Optional start date, inclusive. Requires 'date_field'.
to	Optional end date, exclusive. Requires 'date_field'.
date_field	Optional date or datetime column to use with 'date', 'from', or 'to'. This must be supplied when any date filter is used. Users can identify available date columns by inspecting the dataset on the San Francisco Open Data Portal or by pulling a small sample with 'limit'.
where	Optional raw SoQL 'WHERE' clause for advanced filtering. SoQL is the Socrata Query Language used by San Francisco Open Data. If 'date', 'from', or 'to' are also supplied, their generated conditions are combined with 'where' using 'AND'.
order	Optional raw SoQL 'ORDER BY' clause, such as "DATE_FIELD DESC".
timeout_sec	Request timeout in seconds. Defaults to 30.
clean_names	Logical. If 'TRUE', column names are converted to snake_case using [janitor::clean_names()]. Defaults to 'TRUE'.
coerce_types	Logical. If 'TRUE', the package attempts lightweight, heuristic-based type coercion after downloading the data. Columns are converted only when at least 95 percent of non-missing values can be parsed as the target type. This helps avoid unsafe conversions when source data are inconsistent.

Details

When a catalog ‘key’ is supplied, ‘sf_pull_dataset()’ first retrieves the live San Francisco Open Data catalog to look up the corresponding Socrata ‘dataset_id’, then sends a second request to download the dataset itself. Supplying a ‘dataset_id’ directly is more stable and avoids ambiguity, while keys are provided for readability and classroom-friendly workflows.

Dataset keys are generated from dataset names using [janitor::make_clean_names()]. Because keys are derived from live catalog metadata, Socrata IDs are the most stable identifiers.

‘sf_pull_dataset()’ is designed for common catalog-based workflows. For arbitrary Socrata JSON endpoints that are not included in the package catalog, use [sf_any_dataset()].

The ‘filters’ argument is intended for simple exact-match filtering. For more complex conditions, use the ‘where’ argument with raw SoQL syntax.

Internally, filter field names are wrapped in ‘TRIM()’ when constructing SoQL queries to reduce mismatches caused by leading or trailing whitespace in source data.

Type coercion is intentionally conservative. When ‘coerce_types = TRUE’, the package attempts to infer common R column types from the API response, but columns with inconsistent values may remain character columns.

Datetime coercion is also conservative. Timezone offsets and sub-second precision may not always be preserved during automatic parsing, and columns with inconsistent datetime formats may remain character columns.

Value

A tibble containing rows from the requested San Francisco Open Data dataset.

Examples

```
if (interactive() && curl::has_internet()) {
  # Pull by human-readable key
  sf_pull_dataset("health_inspection_scores_2016_2019", limit = 3)

  # Pull by Socrata ID
  sf_pull_dataset("pyih-qa8i", limit = 3)

  # Filter to one value
  sf_pull_dataset(
    "pyih-qa8i",
    limit = 3,
    filters = list(business_name = "Ike's Kitchen")
  )

  # Filter to multiple values
  sf_pull_dataset(
    "pyih-qa8i",
    limit = 10,
    filters = list(business_name = c("Ike's Kitchen", "Laurel Court"))
  )

  # Date filtering
```

```
sf_pull_dataset(  
  "pyih-qa8i",  
  from = "2018-01-01",  
  to = "2020-01-01",  
  date_field = "inspection_date",  
  limit = 100  
)  
  
# Advanced filtering with raw SoQL  
sf_pull_dataset(  
  "pyih-qa8i",  
  where = "business_name = 'Laurel Court' AND inspection_score = '85'",  
  order = "inspection_date DESC",  
  limit = 100  
)  
}
```

Index

`sf_any_dataset`, [2](#)
`sf_list_datasets`, [3](#)
`sf_pull_dataset`, [4](#)