What's next?

What's next?

- APIs and queries: access OpenData programmatically
- Big data
- Databases
- Web scraping
- Automation
- Data dashboards
- Causal inference

APIs and queries

- APIs : interfaces for getting data from online sources

APIs and queries

- `requests`
  - Tool for running http requests to get data from (or send data to) online resources
  - *https://realpython.com/python-requests/*

Databases

- SQL
  - Standard query language for getting data from databases
  - Intro tutorial: *https://sqlbolt.com*

SQL

```sql
SELECT
    `station_complex_id`,
    `station_complex`,
    `payment_method`,
    sum(`ridership`) AS `sum_ridership`
WHERE
    `transit_timestamp`
    BETWEEN "2023-01-01T00:00:00"
    AND "2024-12-31T23:59:59"
GROUP BY
    `station_complex_id`,
    `station_complex`,
    `payment_method`
```

NYC OpenData requests

NYC OpenData

Home   Data   About ⌄   Learn ⌄   Contact Us   🔍 | Sign In

← Back to overview

Export dataset   ✕

Only the data returned by your current query will be exported.

Download file   **API endpoint**

Data format
JSON   ⌄

API endpoint
https://data.cityofnewyork.us/resource/erm2-nwe9.jso

Cancel   Copy to clipboard

✕   Export

📅 by_month_created_date
   by_month_created_date

≡

2010 Jan 01 12:00:00 AM

2010 Jan 01 12:00:00 AM

2010 Jan 01 12:00:00 AM

2010 Jan 01 12:00:00 AM

2010 Jan 01 12:00:00 AM

‹ | 1 | of 36 ›   *Showing rows 1-100 of 3564*

▼   Group & aggregate | ✕ Clear all

Apply   Success!

Sign In to NYC Open Data

✉ Email   🔑⌄

🔒 Password

Forgot Password?

Sign In

Don't have an account yet? Sign Up

Home   Data   About ⌄   Learn ⌄   Contact Us   🔍   🔔   Dan

Developer Settings

API Keys

▶ What are API Keys? ❓

Search

| 🔍 Name or ID | Include Deleted ☐ | Search |

Create new API Key

Api Key Name	Api Key ID	Last Used At	Created At ⌄	Actions
No Results				

Previous   Next

App Tokens

▶ What are App Tokens? ❓

Create New App Token

Name	Description	App Token	Secret Token	Actions

Developer Settings

API Keys

▶ What are API Keys? ❓

Search

| 🔍 Name or ID | Include Deleted ☐ | Search | | Create new API Key |

Api Key Name	Api Key ID	Last Used At	Created At ↓	Actions
No Results				

| Previous | Next |

App Tokens

▶ What are App Tokens? ❓

Create New App Token

Name	Description	App Token	Secret Token	Actions
open-data-app-token-demo	demo app token	ap0EJq7UmZvfXcVl6cBkFCB1G 📋	Show Secret Token	•••

NYC OpenData

Home   Data   About ⌄   Learn ⌄   Contact Us   🔍 | Sign In

← Back to overview

| 📅 | by_month_created_date |
| | *by_month_created_date* |

2010 Jan 01 12:00:00 AM

2010 Jan 01 12:00:00 AM

2010 Jan 01 12:00:00 AM

2010 Jan 01 12:00:00 AM

2010 Jan 01 12:00:00 AM

< | 1 | of 36 | >

Group & aggregate | ⊗ Clear all

Apply   Success!

Export dataset ✕

Only the data returned by your current query will be exported.

Download file   **API endpoint**

Data format
JSON   ⌄

API endpoint
https://data.cityofnewyork.us/resource/erm2-nwe9.jso   📋

Cancel   **Copy to clipboard**

✕   **Export**

≡

Showing rows 1-100 of 3564

https://data.cityofnewyork.us/resource/erm2-nwe9.json?$query=SELECT%
0A%20%20date_trunc_ym(%60created_date%60)%20AS%20%60by_month_created
_date%60%2C%0A%20%20%60complaint_type%60%2C%0A%20%20count(DISTINCT%2
0%60unique_key%60)%20AS%20%60count_distinct_unique_key%60%0AWHERE%20
%60agency%60%20%3D%3D%20%22DSNY%22%0AGROUP%20BY%20date_trunc_ym(%60c
reated_date%60)%2C%20%60complaint_type%60

```
https://data.cityofnewyork.us/resource/erm2-nwe9.json?$query=
SELECT
  date_trunc_ym(`created_date`) AS `by_month_created_date`,
  `complaint_type`,
  count(DISTINCT `unique_key`) AS `count_distinct_unique_key`
WHERE `agency` == "DSNY"
GROUP BY date_trunc_ym(`created_date`), `complaint_type`
```

```
https://data.cityofnewyork.us/resource/erm2-nwe9.json?$query=
SELECT
    date_trunc_ym(`created_date`) AS `by_month_created_date`,
    `complaint_type`,
    count(DISTINCT `unique_key`) AS `count_distinct_unique_key`
WHERE `agency` == "DSNY"
GROUP BY date_trunc_ym(`created_date`), `complaint_type`
LIMIT 1000000000
```

```
https://data.cityofnewyork.us/resource/erm2-nwe9.json?$query=
SELECT
  date_trunc_ym(`created_date`) AS `by_month_created_date`,
  `complaint_type`,
  count(DISTINCT `unique_key`) AS `count_distinct_unique_key`
WHERE `agency` == "DSNY"
GROUP BY date_trunc_ym(`created_date`), `complaint_type`
LIMIT 1000000000
&$$app_token=xxxxxxxxxxxxxxx
```

https://data.cityofnewyork.us/resource/erm2-nwe9.json?$query=SELECT%0A%20%20date_trunc_ym(%60created_date%60)%20AS%20%60by_month_created_date%60%2C%0A%20%20%60complaint_type%60%2C%0A%20%20count(DISTINCT%20%60unique_key%60)%20AS%20%60count_distinct_unique_key%60%0AWHERE%20%60agency%60%20%3D%3D%20%22DSNY%22%0AGROUP%20BY%20date_trunc_ym(%60created_date%60)%2C%20%60complaint_type%60%20LIMIT%201000000000&$$app_token=xxxxxxxxxxxxxxx

```python
OPEN_DATA_APP_TOKEN = 'xxx'
BASE_URL = 'https://data.cityofnewyork.us/'

table_id = 'erm2-nwe9'

query = """
    SELECT
        DATE_TRUNC_YMD(`created_date`) as date,
        COUNT(`unique_key`) as count
    WHERE
        (`complaint_type` LIKE "Homeless Person Assistance")
        AND (`created_date` >= "2019-01-01")
    GROUP BY
        DATE_TRUNC_YMD(`created_date`)
    LIMIT 100000
    """

data_request_url = (
    BASE_URL
    + 'resource/'
    + table_id
    +'.json?'
    + '$query='
    + urllib.parse.quote(query)
    + '&$$app_token='
    + OPEN_DATA_APP_TOKEN
)

data = pd.read_json(data_request_url)
```

Big data

- dask
  - pandas syntax on big (including bigger-than-memory) datasets
  - *https://www.dask.org*
- DuckDB
  - Builds standalone database to run queries with SQL in Python
  - Handles bigger-than-memory data and loads and queries *fast*
  - *https://duckdb.org/docs/stable/clients/python/overview.html*
  - *https://realpython.com/python-duckdb/*
- polars
  - Alternative dataframe syntax with *fast* execution on big data
  - *https://docs.pola.rs/#key-features*

Web scraping

- BeautifulSoup
  - For parsing web pages and extracting links, text, or data tables
  - *https://beautiful-soup-4.readthedocs.io/en/latest/*
  - *https://realpython.com/beautiful-soup-web-scraper-python/*

Automation

- Github Actions
  - For running tasks on a schedule
  - e.g. saving data from real-time Citi Bike stations feed: *https://github.com/NYCComptroller/citi-bike-gbfs*
  - *https://palewi.re/docs/first-github-scraper/action.html*
  - *https://githubnext.com/projects/flat-data*

Data dashboards

- Streamlit
  - *https://streamlit.io/#install*
- Dash
  - *https://dash.plotly.com/tutorial*
  - *https://realpython.com/python-dash/*
- Observable
  - *https://observablehq.com/framework/*

Causal inference

- Using statistics to determine if something caused something else
  - Causal Inference for the Brave and True
    https://matheusfacure.github.io/python-causality-handbook/

Where to learn

- Kaggle
  - *https://www.kaggle.com/*
- Real Python
  - *https://realpython.com/*
- Think Python
  - *https://allendowney.github.io/ThinkPython*
- BetaNYC
  - *https://www.beta.nyc/featured-tools/nyc-opendata-classes/*