# PANDAS DATAFRAME

GLY606 Water Data Analysis & Modeling

Sep 20th 2024


University at Buffalo The State University of New York

# Homework #2 Done!

In the last week, we introduced **Numpy** (a powerful tool to generate data arrays and calculations)

How can we more effectively **manipulate** data?

```
import pandas as pd
```

Data Structure

Powerful functions

# Data Structure

| Data Structure | Dimension |
|---|---|
| Series | 1 |
| Data Frames | 2 |

**What is the difference between series and Data Frames?**

Data Series

```
tom      105
bob      306
nancy   3560
dan     1200
eric      50
```

Data Framework

| | Fav_number | Fav_color |
|---|---|---|
| **tom** | 105 | red |
| **bob** | 306 | blue |
| **nancy** | 3560 | orange |
| **dan** | 1200 | pink |
| **eric** | 50 | green |

# DataFrame syntax

column

Row

| | Fav_number | Fav_color |
|---|---|---|
| **tom** | 105 | red |
| **bob** | 306 | blue |
| **nancy** | 3560 | orange |
| **dan** | 1200 | pink |
| **eric** | 50 | green |

# DataFrame syntax

Column name

index

| | Fav_number | Fav_color |
|---|---|---|
| **tom** | 105 | red |
| **bob** | 306 | blue |
| **nancy** | 3560 | orange |
| **dan** | 1200 | pink |
| **eric** | 50 | green |

# DataFrame syntax

**df.loc[index, column name]**

Column name

index

| | Fav_number | Fav_color |
|---|---|---|
| **tom** | 105 | red |
| **bob** | 306 | blue |
| **nancy** | 3560 | orange |
| **dan** | 1200 | pink |
| **eric** | 50 | green |

# DataFrame syntax

**df.loc['tom','Fav_number']**

Column name

index

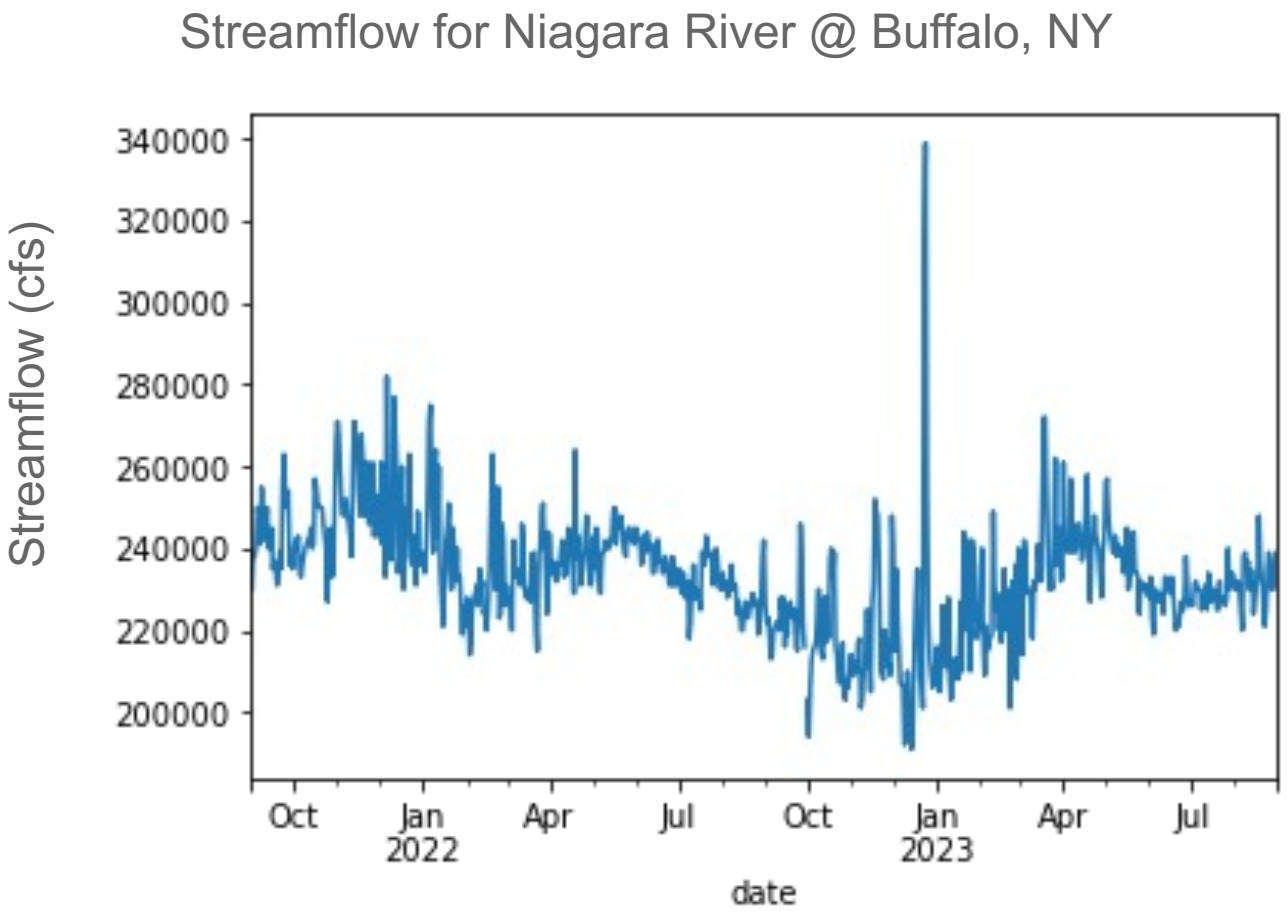
| | Fav_number | Fav_color |
|---|---|---|
| **tom** | 105 | red |
| **bob** | 306 | blue |
| **nancy** | 3560 | orange |
| **dan** | 1200 | pink |
| **eric** | 50 | green |

# DataFrame syntax

**df.iloc[0,0]**

| | 0th column | 1st column |
|---|---|---|
| | **Fav_number** | **Fav_color** |
| 0th row **tom** | 105 | red |
| 1st row **bob** | 306 | blue |
| 2nd row **nancy** | 3560 | orange |
| 3rd row **dan** | 1200 | pink |
| 4th row **eric** | 50 | green |

# Data manipulation for time series data



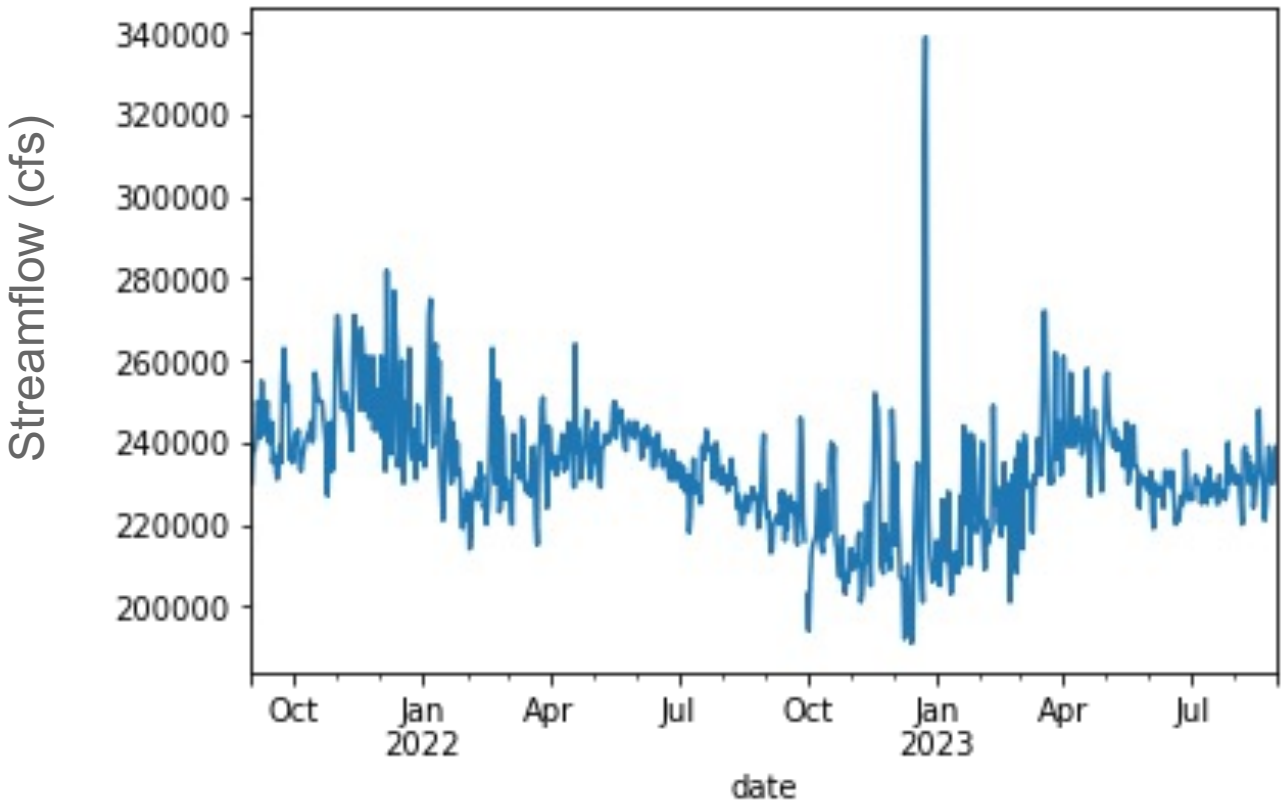

Out[70]:

| date | streamflow | quality_flag |
|---|---|---|
| **2022-10-01** | 203000.0 | A |
| **2022-10-02** | 194000.0 | A |
| **2022-10-03** | 206000.0 | A |
| **2022-10-04** | 213000.0 | A |
| **2022-10-05** | 215000.0 | A |
| **...** | ... | ... |
| **2022-12-28** | 209000.0 | A |
| **2022-12-29** | 206000.0 | A |
| **2022-12-30** | 207000.0 | A |
| **2022-12-31** | 211000.0 | A |
| **2023-01-01** | 216000.0 | A |

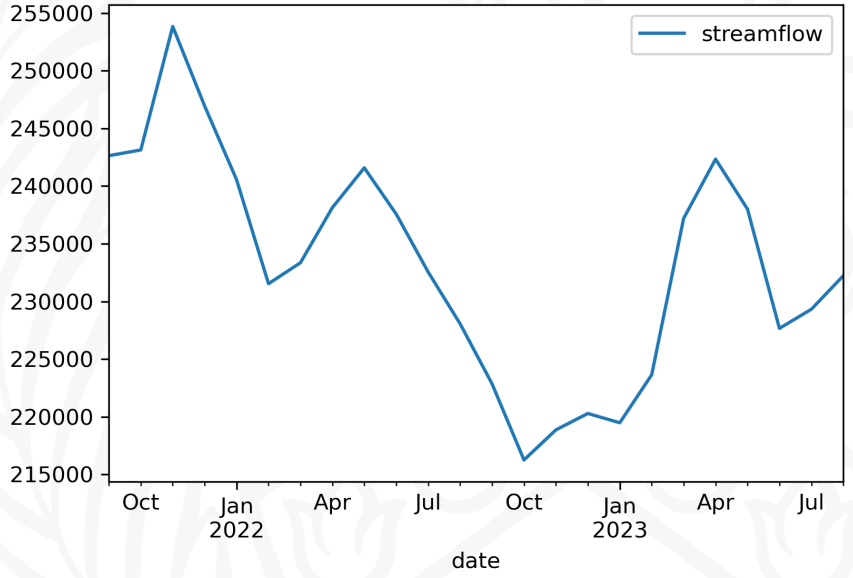
93 rows × 2 columns

# Data manipulation for time series data

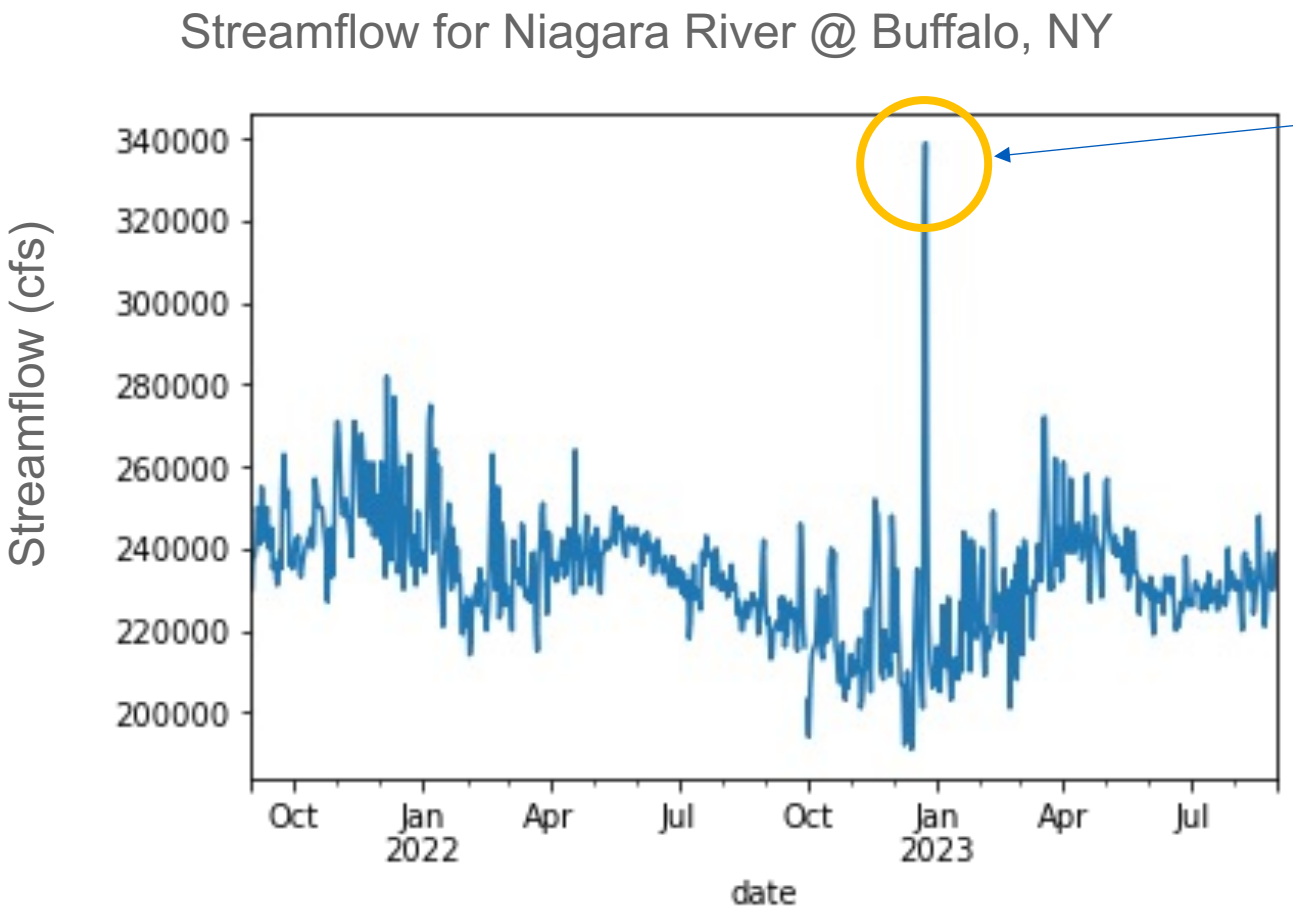
Streamflow for Niagara River @ Buffalo, NY




How can we change the frequency of data from daily to monthly?



# Data manipulation for time series data




How can we identify the extreme high flow events?

When is that event?

# These can all be achieved in Pandas DataFrame!

- https://github.com/act-hydro/GLY606_2024/blob/main/in_class_practice/python_practice/python_inclass_5_dataframe.ipynb

- Data
  - https://github.com/act-hydro/GLY606_2024/blob/main/in_class_practice/python_practice/flow_cfs.USGS_04216000.Niagara_river.csv

# Homework #3 is coming!

- It will focus on Python Numpy & Matplotlib.
- Due Date: **1pm, Sep 27th 2024 (Friday)**
- Submission: Save the notebook as a PDF and turn in the PDF
- Platform: UBLearns (preferred) or email