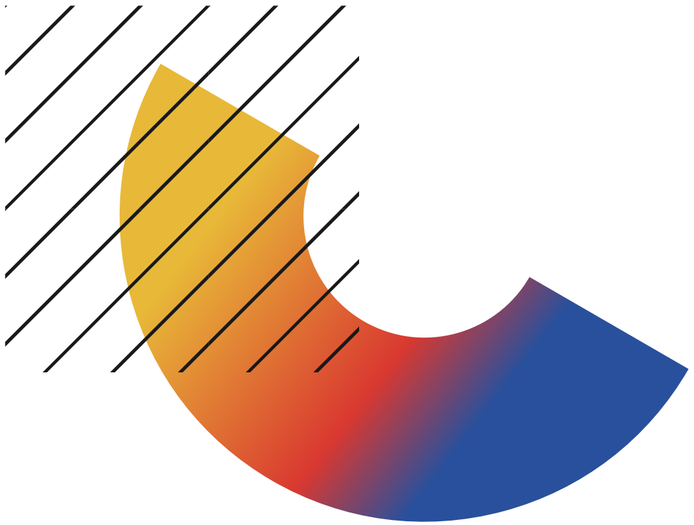
IST652 SCRIPTING FOR DATA ANALYSIS

Tokyo Olympics 2021: A Python Analysis

Farahin Choudhury

Background



- Datasets from Kaggle
- Set of 5 excel spreadsheets (.xlsx)
- Athletes, gender, teams (countries), coaches, medals
- Varying number of observations
- Data types standardized

Data questions

1. Which athlete received the highest number of medals?
2. What are the top and bottom 3 sports based on medal count?
3. Which country among the biggest delegation had the highest medal count?
4. Which country among the smallest delegation had the highest medal count?
5. How many past Olympic events have the athletes competed in?

Main libraries

pandas

Main library for the analysis.

numpy

Important for doing calculations

CSV

File conversion from .xlsx to .csv (familiarity)

Data cleaning and preparation

ways the data needed to be prepared for analysis

01 Import csv and pandas modules

02 Convert excel files to csv files

03 Use pandas to create dataframes to explore data

04 Drop and rename some columns

05 Merge all 5 data frames into one composite data frame called OlympicsData

INTERESTING FINDINGS

Example data frame

```
: 1 medals = medals.drop(['Rank'], axis = 1)
  2 medals.columns =['Olympic Committee Name', 'Gold', 'Silver', 'Bronze', 'Total Medals', 'Rank by Total']
```

```
: 1 medals = medals.set_index('Olympic Committee Name')
```

```
: 1 medals
```

:

Olympic Committee Name	Gold	Silver	Bronze	Total Medals	Rank by Total
United States of America	39	41	33	113	1
People's Republic of China	38	32	18	88	2
Japan	27	14	17	58	5
Great Britain	22	21	22	65	4
ROC	20	28	23	71	3
...	...	...	...	...	...
Ghana	0	0	1	1	77
Grenada	0	0	1	1	77
Kuwait	0	0	1	1	77
Republic of Moldova	0	0	1	1	77
Syrian Arab Republic	0	0	1	1	77

93 rows × 5 columns

Possible avenues for further analysis

API
JSON + MongoDB
Machine Learning

Thank you!