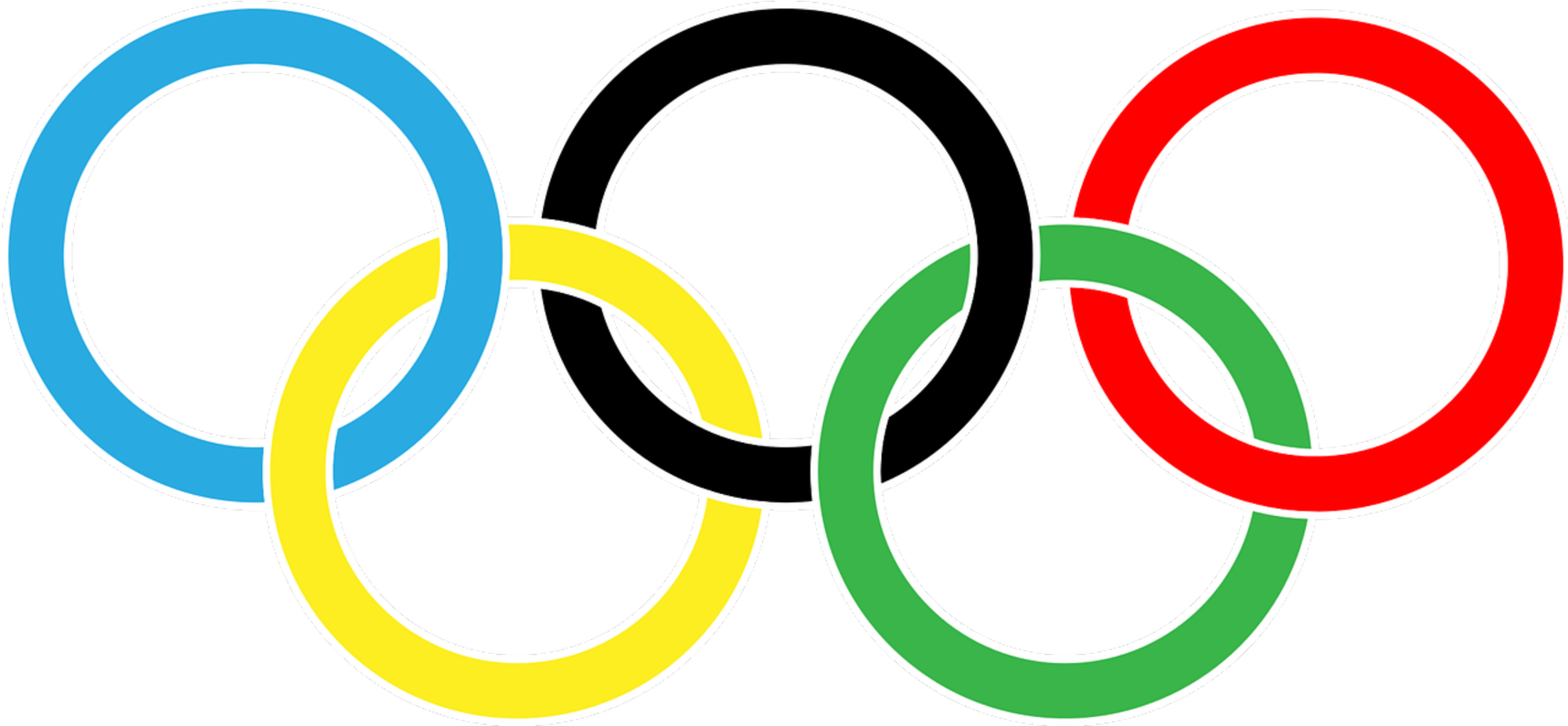
MERGING DATAFRAMES WITH PANDAS

Medals in the Summer Olympics

Does a host country win more medals?



Summer Olympic medalists 1896 to 2008 - IOC COUNTRY CODES.csv

Country	NOC	ISO code
Afghanistan	AFG	AF
Albania	ALB	AL
Algeria	ALG	DZ
American Samoa*	ASA	AS
Andorra	AND	AD
Angola	ANG	AO
Antigua and Barbuda	ANT	AG
Argentina	ARG	AR
Armenia	ARM	AM
Aruba*	ARU	AW
Australia	AUS	AU
Austria	AUT	AT

Summer Olympic medalists 1896 to 2008 - EDITIONS.tsv

Edition	Bronze	Gold	Silver	Grand Total	City	Country
1896	40	64	47	151	Athens	Greece
1900	142	178	192	512	Paris	France
1904	123	188	159	470	St. Louis	United States
1908	211	311	282	804	London	United Kingdom
1912	284	301	300	885	Stockholm	Sweden
1920	355	497	446	1298	Antwerp	Belgium
1924	285	301	298	884	Paris	France
1928	242	229	239	710	Amsterdam	Netherlands
1932	196	213	206	615	Los Angeles	United States
1936	282	299	294	875	Berlin	Germany
1948	268	276	270	814	London	United Kingdom
1952	299	300	290	889	Helsinki	Finland

summer_1896.csv, summer_1900.csv, ..., summer_2008.csv

Sport	Discipline	Athlete	NOC	Gender	Event	Event_gender	Medal
Aquatics	Diving	XIAO, Hailiang	CHN	Men	10m platform	M	Bronze
Aquatics	Diving	SAUTIN, Dmitry	RUS	Men	10m platform	M	Gold
Aquatics	Diving	HEMPEL, Jan	GER	Men	10m platform	M	Silver
Aquatics	Diving	CLARK, Mary Ellen	USA	Women	10m platform	W	Bronze
Aquatics	Diving	FU, Mingxia	CHN	Women	10m platform	W	Gold
Aquatics	Diving	WALTER, Annika	GER	Women	10m platform	W	Silver
Aquatics	Diving	LENZI, Mark Edward	USA	Men	3m springboard	M	Bronze
Aquatics	Diving	XIONG, Ni	CHN	Men	3m springboard	M	Gold
Aquatics	Diving	YU, Zhuocheng	CHN	Men	3m springboard	M	Silver
Aquatics	Diving	PELLETIER, Annie	CAN	Women	3m springboard	W	Bronze
Aquatics	Diving	FU, Mingxia	CHN	Women	3m springboard	W	Gold
Aquatics	Diving	LASHKO, Irina	RUS	Women	3m springboard	W	Silver

Reminder: loading & merging files

- pd.read_csv() (& its many options)
- Looping over files, e.g.,
  - [pd.read_csv(f) for f in glob('*.csv')]
- Concatenating & appending, e.g.,
  - pd.concat([df1, df2], axis=0)
  - df1.append(df2)

MERGING DATAFRAMES WITH PANDAS

Let’s practice!

MERGING DATAFRAMES WITH PANDAS

Quantifying performance

Medals DataFrame

	Sport	Discipline	Athlete	NOC	Gender	Event	Event_gender	Medal	Edition
0	Aquatics	Swimming	HAJOS, Alfred	HUN	Men	100m freestyle	M	Gold	1896
1	Aquatics	Swimming	HERSCHMANN, Otto	AUT	Men	100m freestyle	M	Silver	1896
2	Aquatics	Swimming	DRIVAS, Dimitrios	GRE	Men	100m freestyle for sailors	M	Bronze	1896
3	Aquatics	Swimming	MALOKINIS, Ioannis	GRE	Men	100m freestyle for sailors	M	Gold	1896
4	Aquatics	Swimming	CHASAPIS, Spiridon	GRE	Men	100m freestyle for sailors	M	Silver	1896
5	Aquatics	Swimming	CHOROPHAS, Efstathios	GRE	Men	1200m freestyle	M	Bronze	1896
6	Aquatics	Swimming	HAJOS, Alfred	HUN	Men	1200m freestyle	M	Gold	1896
7	Aquatics	Swimming	ANDREOU, Joannis	GRE	Men	1200m freestyle	M	Silver	1896
8	Aquatics	Swimming	CHOROPHAS, Efstathios	GRE	Men	400m freestyle	M	Bronze	1896
9	Aquatics	Swimming	NEUMANN, Paul	AUT	Men	400m freestyle	M	Gold	1896
10	Aquatics	Swimming	PEPANOS, Antonios	GRE	Men	400m freestyle	M	Silver	1896
11	Athletics	Athletics	LANE, Francis	USA	Men	100m	M	Bronze	1896

Constructing a pivot table

- Apply DataFrame pivot_table() method
  - index: column to use as index of pivot table
  - values: column(s) to aggregate
  - aggfunc: function to apply for aggregation
  - columns: categories as columns of pivot table

Constructing a pivot table

NOC	AFG	AHO	ALG	ANZ	ARG	ARM	AUS	AUT	AZE	BAH	...	URS	URU	USA	UZB	VEN	VIE	YUG	ZAM	ZIM	ZZX
Edition																					
1896	NaN	NaN	NaN	NaN	NaN	NaN	2.0	5.0	NaN	NaN	...	NaN	NaN	20.0	NaN	NaN	NaN	NaN	NaN	NaN	6.0
1900	NaN	NaN	NaN	NaN	NaN	NaN	5.0	6.0	NaN	NaN	...	NaN	NaN	55.0	NaN	NaN	NaN	NaN	NaN	NaN	34.0
1904	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	...	NaN	NaN	394.0	NaN	NaN	NaN	NaN	NaN	NaN	8.0
1908	NaN	NaN	NaN	19.0	NaN	NaN	NaN	1.0	NaN	NaN	...	NaN	NaN	63.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1912	NaN	NaN	NaN	10.0	NaN	NaN	NaN	14.0	NaN	NaN	...	NaN	NaN	101.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1920	NaN	NaN	NaN	NaN	NaN	NaN	6.0	NaN	NaN	NaN	...	NaN	NaN	193.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1924	NaN	NaN	NaN	NaN	11.0	NaN	10.0	4.0	NaN	NaN	...	NaN	22.0	198.0	NaN	NaN	NaN	2.0	NaN	NaN	NaN
1928	NaN	NaN	NaN	NaN	32.0	NaN	4.0	4.0	NaN	NaN	...	NaN	22.0	84.0	NaN	NaN	NaN	12.0	NaN	NaN	NaN
1932	NaN	NaN	NaN	NaN	4.0	NaN	5.0	5.0	NaN	NaN	...	NaN	1.0	181.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1936	NaN	NaN	NaN	NaN	11.0	NaN	1.0	50.0	NaN	NaN	...	NaN	NaN	92.0	NaN	NaN	NaN	1.0	NaN	NaN	NaN
1948	NaN	NaN	NaN	NaN	12.0	NaN	16.0	4.0	NaN	NaN	...	NaN	3.0	148.0	NaN	NaN	NaN	16.0	NaN	NaN	NaN
1952	NaN	NaN	NaN	NaN	6.0	NaN	20.0	3.0	NaN	NaN	...	117.0	14.0	130.0	NaN	1.0	NaN	24.0	NaN	NaN	NaN

Computing fractions

NOC	AFG	AHO	ALG	ANZ	ARG	ARM	AUS	AUT	AZE	BAH	...	URS	URU	USA	UZB
Edition															
1896	NaN	NaN	NaN	NaN	NaN	NaN	0.013245	0.033113	NaN	NaN	...	NaN	NaN	0.132450	NaN
1900	NaN	NaN	NaN	NaN	NaN	NaN	0.009766	0.011719	NaN	NaN	...	NaN	NaN	0.107422	NaN
1904	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.002128	NaN	NaN	...	NaN	NaN	0.838298	NaN
1908	NaN	NaN	NaN	0.023632	NaN	NaN	NaN	0.001244	NaN	NaN	...	NaN	NaN	0.078358	NaN
1912	NaN	NaN	NaN	0.011299	NaN	NaN	NaN	0.015819	NaN	NaN	...	NaN	NaN	0.114124	NaN
1920	NaN	NaN	NaN	NaN	NaN	NaN	0.004622	NaN	NaN	NaN	...	NaN	NaN	0.148690	NaN
1924	NaN	NaN	NaN	NaN	0.012443	NaN	0.011312	0.004525	NaN	NaN	...	NaN	0.024887	0.223982	NaN
1928	NaN	NaN	NaN	NaN	0.045070	NaN	0.005634	0.005634	NaN	NaN	...	NaN	0.030986	0.118310	NaN
1932	NaN	NaN	NaN	NaN	0.006504	NaN	0.008130	0.008130	NaN	NaN	...	NaN	0.001626	0.294309	NaN
1936	NaN	NaN	NaN	NaN	0.012571	NaN	0.001143	0.057143	NaN	NaN	...	NaN	NaN	0.105143	NaN

MERGING DATAFRAMES WITH PANDAS

Let’s practice!

MERGING DATAFRAMES WITH PANDAS

Reshaping and plotting

Reshaping the data

NOC	AFG	AHO	ALG	ANZ	ARG	ARM	...	VEN	VIE	YUG	ZAM	ZIM	ZZX
Edition													
1896	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN
1900	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	33.561198
1904	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	-22.642384
1908	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	0.000000
1912	NaN	NaN	NaN	-26.092774	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	0.000000
1920	NaN	NaN	NaN	0.000000	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	0.000000
1924	NaN	NaN	NaN	0.000000	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	0.000000
1928	NaN	NaN	NaN	0.000000	131.101152	NaN	...	NaN	NaN	323.521127	NaN	NaN	0.000000
1932	NaN	NaN	NaN	0.000000	-25.794206	NaN	...	NaN	NaN	0.000000	NaN	NaN	0.000000
1936	NaN	NaN	NaN	0.000000	-10.271982	NaN	...	NaN	NaN	-29.357594	NaN	NaN	0.000000
1948	NaN	NaN	NaN	0.000000	-4.601500	NaN	...	NaN	NaN	47.596769	NaN	NaN	0.000000
1952	NaN	NaN	NaN	0.000000	-10.508545	NaN	...	NaN	NaN	34.043608	NaN	NaN	0.000000

	Edition	NOC	Change
0	1896	AFG	NaN
1	1900	AFG	NaN
2	1904	AFG	NaN
3	1908	AFG	NaN
4	1912	AFG	NaN
5	1920	AFG	NaN
6	1924	AFG	NaN
7	1928	AFG	NaN
8	1932	AFG	NaN
9	1936	AFG	NaN
10	1948	AFG	NaN
11	1952	AFG	NaN

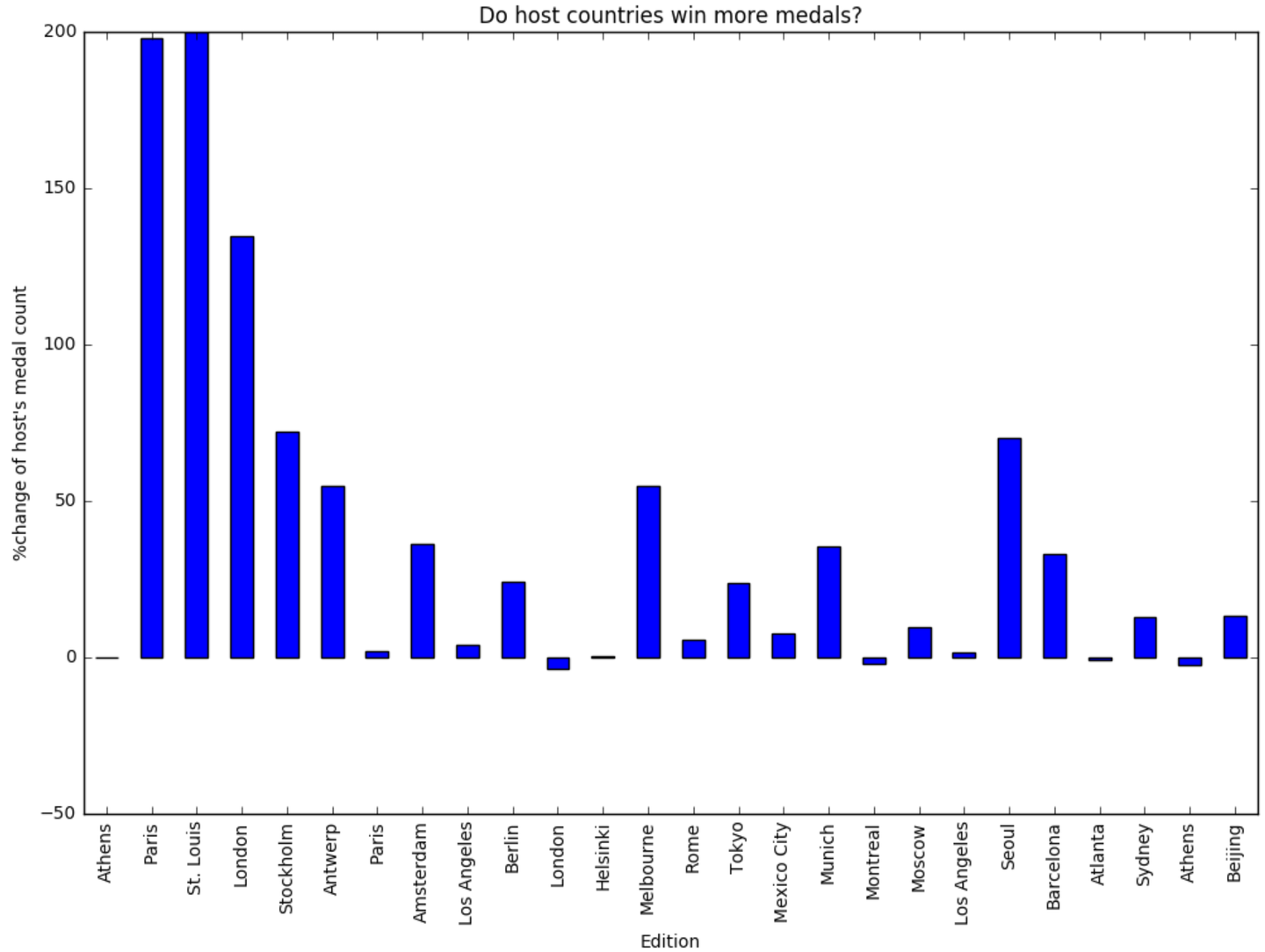
Host country data

	Edition	Bronze	Gold	Silver	Grand Total	City	Country	Host_NOC
0	1896	40	64	47	151	Athens	Greece	GRE
1	1900	142	178	192	512	Paris	France	FRA
2	1904	123	188	159	470	St. Louis	United States	USA
3	1908	211	311	282	804	London	United Kingdom	GBR
4	1912	284	301	300	885	Stockholm	Sweden	SWE
5	1920	355	497	446	1298	Antwerp	Belgium	BEL
6	1924	285	301	298	884	Paris	France	FRA
7	1928	242	229	239	710	Amsterdam	Netherlands	NED
8	1932	196	213	206	615	Los Angeles	United States	USA
9	1936	282	299	294	875	Berlin	Germany	GER
10	1948	268	276	270	814	London	United Kingdom	GBR
11	1952	299	300	290	889	Helsinki	Finland	FIN

Quantifying influence

	NOC	count	Host_NOC	Grand Total	fraction	change
Edition						
1896	GRE	52	GRE	151	0.344371	NaN
1900	FRA	185	FRA	512	0.361328	198.002486
1904	USA	394	USA	470	0.838298	199.651245
1908	GBR	347	GBR	804	0.431592	134.489218
1912	SWE	173	SWE	885	0.195480	71.896226
1920	BEL	188	BEL	1298	0.144838	54.757887
1924	FRA	122	FRA	884	0.138009	2.046362
1928	NED	65	NED	710	0.091549	36.315243
1932	USA	181	USA	615	0.294309	3.739184
1936	GER	210	GER	875	0.240000	24.108011
1948	GBR	56	GBR	814	0.068796	-3.635059
1952	FIN	40	FIN	889	0.044994	0.121662

Graphical summary



MERGING DATAFRAMES WITH PANDAS

Let’s practice!

Final thoughts