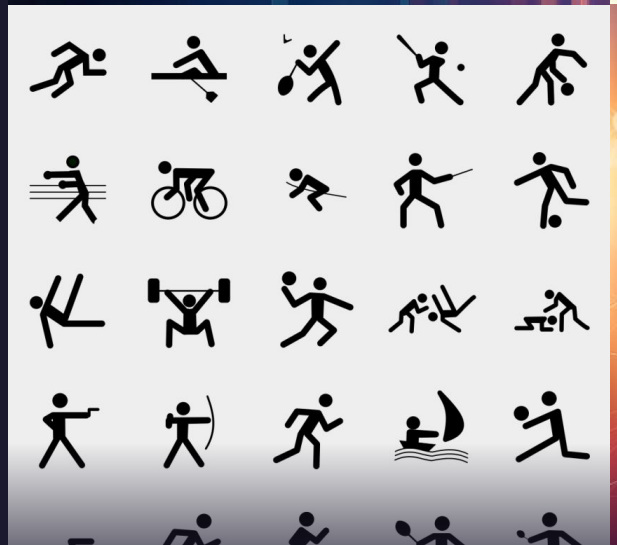
Measurable Factors Influencing Country's success in Olympics

Presented by:

Anam Khalid

Introduction

- According to nbc sports, USA, Russia, Germany, Great Britain and France are the top performers in Olympics. We used various methods to find out the measurable factors that contributes to the success of a country in Olympic games!



Top 10 Success Factors

1. Dedication and Persistence
2. Support of Family and Friends
3. Excellent Coaches
4. Love of sport
5. Excellent Training Programs and Facilities
6. Natural Talent
7. Competitiveness
8. Focus
9. Work Ethic
10. Financial Support

Top 10 Obstacles

1. Lack of Financial Support
2. Conflict with Roles in Life
3. Lack of Coaching Expertise or Support
4. Lack of Support from USOC and NGB
5. Mental Obstacles
6. Lack of Training/Competition Opportunities
7. Medical Problems
8. Lack of Social Support
9. Physical Limitations
10. Failure

Agenda:

- How does population size affect country's success in winning the Olympic Games?
- What is the relationship between GDP and Olympic medals?

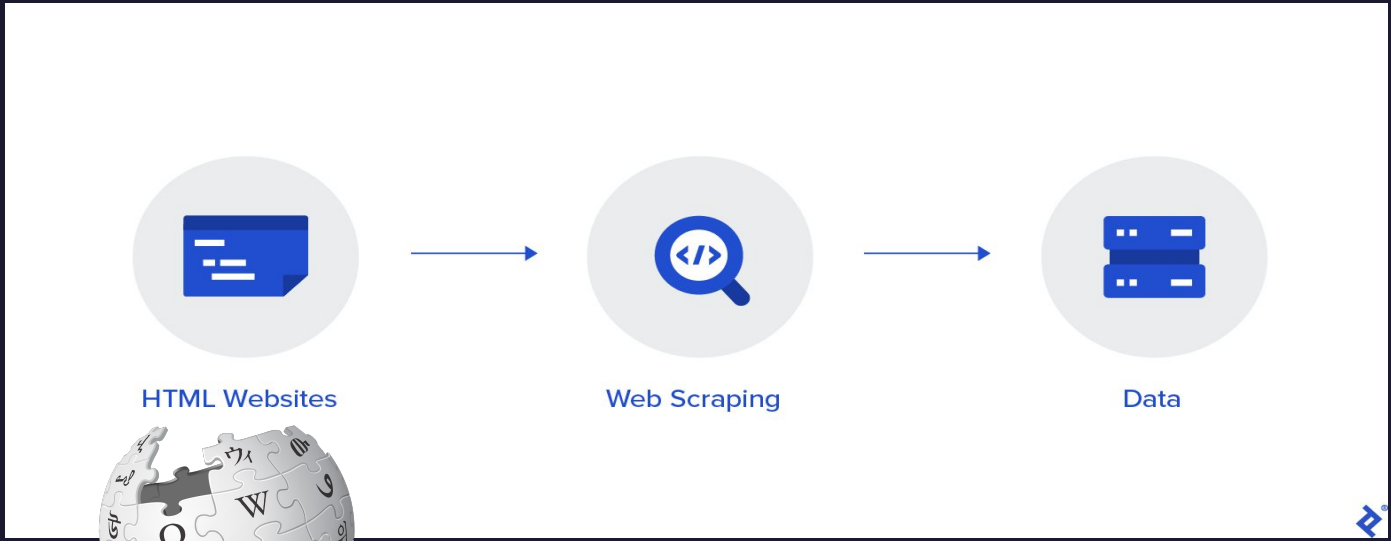
Hypothesis

*I believed that countries with larger **GDP** and **population** would have more success in the Olympic games.*

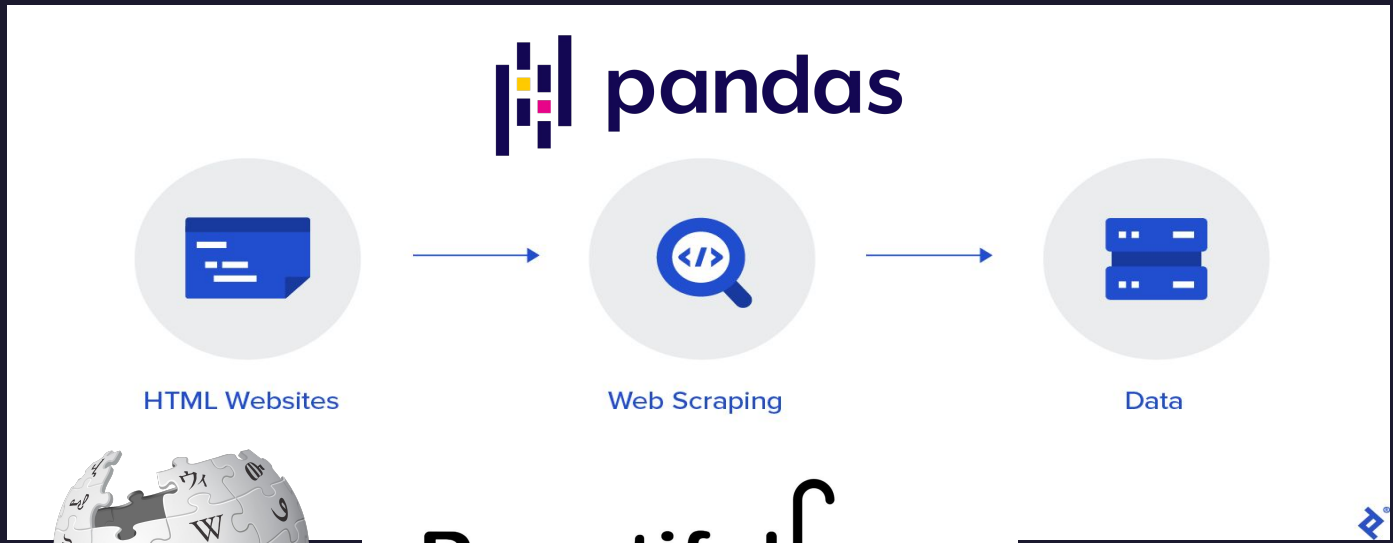
Where did I get the Data from?

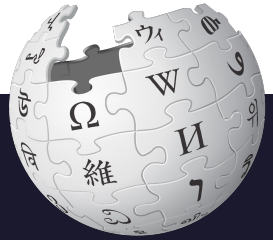
Web Scraping - Where did I get the Data from?

What was the process?



What was the process?





Beautiful Soup

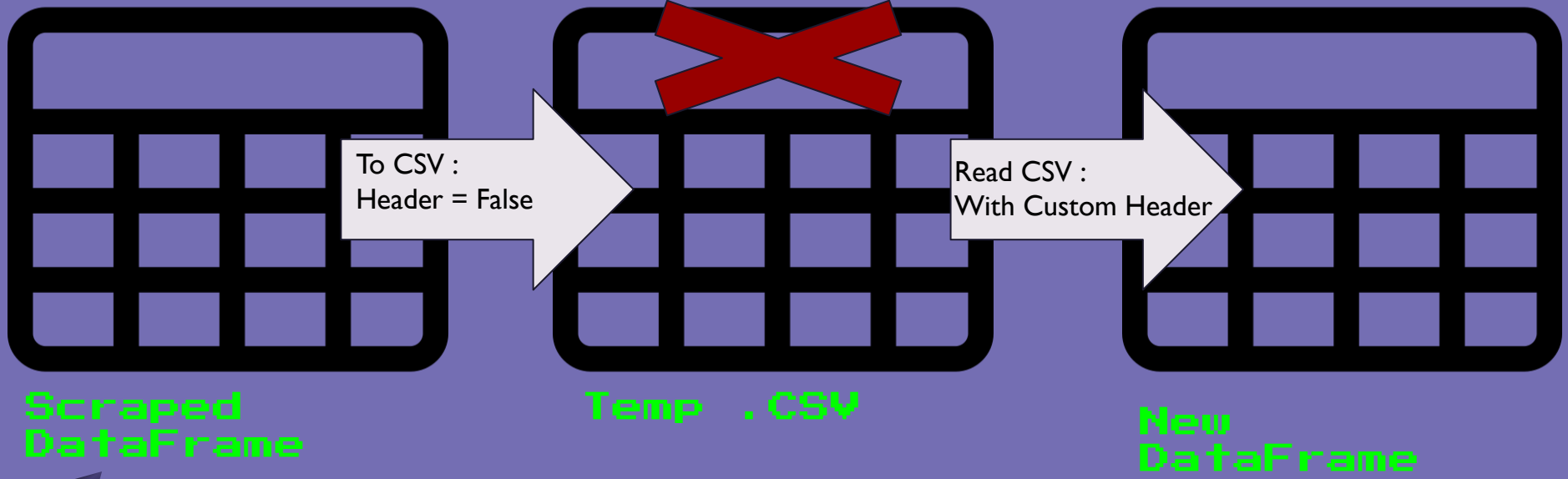
Cleaning the data - What are the problems?

Team	Summer Olympic Games					Winter Olympic Games					Combined total				
Team (IOC code)	No.	1	2	3	Total	No.	1	2	3	Total	No.	1	2	3	Total
Afghanistan (AFG)	15	0	0	2	2	0	0	0	0	0	15	0	0	2	2

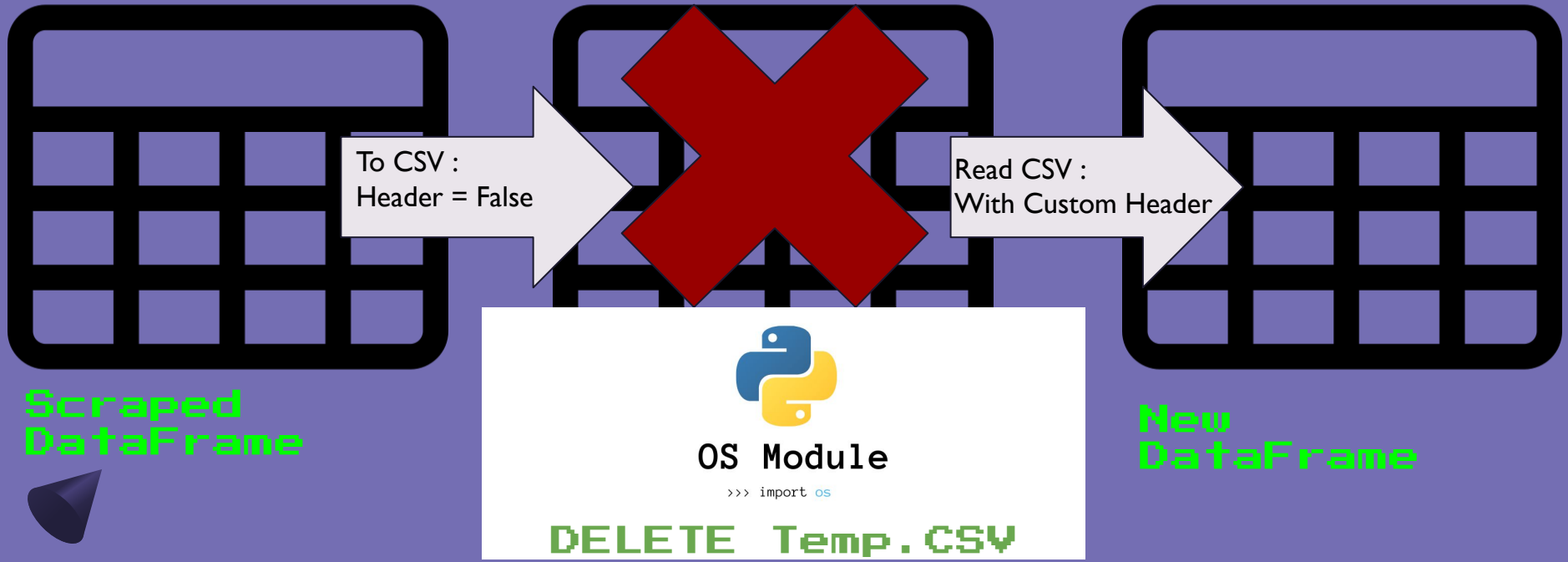
	Team	Summer Olympic Games					Winter Olympic Games					Combined total				
	.mw-parser-output .tooltip-dotted{border-bottom:1px dotted;cursor:help}Team (IOC code)	No.	Unnamed: 2_level_1	Unnamed: 3_level_1	Unnamed: 4_level_1	Total	No.	Unnamed: 7_level_1	Unnamed: 8_level_1	Unnamed: 9_level_1	Total	No.	Unnamed: 12_level_1	Unnamed: 13_level_1	Unnamed: 14_level_1	Tot
0	Afghanistan (AFG)	15	0	0	2	2	0	0	0	0	0	15	0	0	2	

- The two headers that don't make sense
- Too many columns
- Country Codes

The Headers

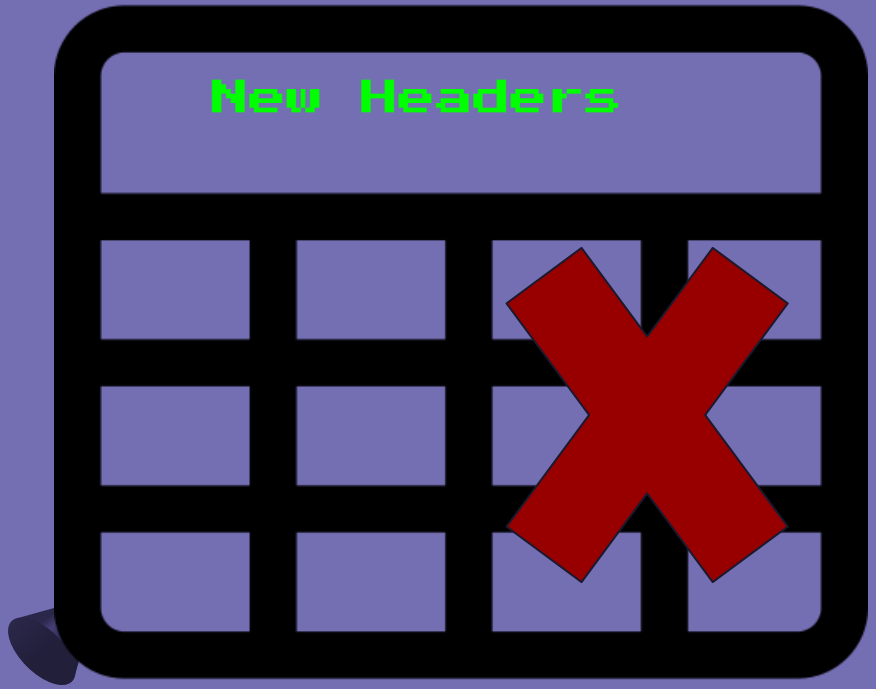

To CSV :
Header = False
Read CSV :
With Custom Header
Scraped DataFrame
Temp .CSV
New DataFrame

The Headers


To CSV :
Header = False
Read CSV :
With Custom Header
Scraped DataFrame
OS Module
>>> import os
DELETE Temp.CSV
New DataFrame

Finishing the cleaning

New DataFrame



country
Afghanistan (AFG)
Algeria (ALG)
Argentina (ARG)
Armenia (ARM)
Australasia (ANZ) [ANZ]
...
Zimbabwe (ZIM) [ZIM]

Delete "(" and everything to the Right

country
Afghanistan
Algeria
Argentina

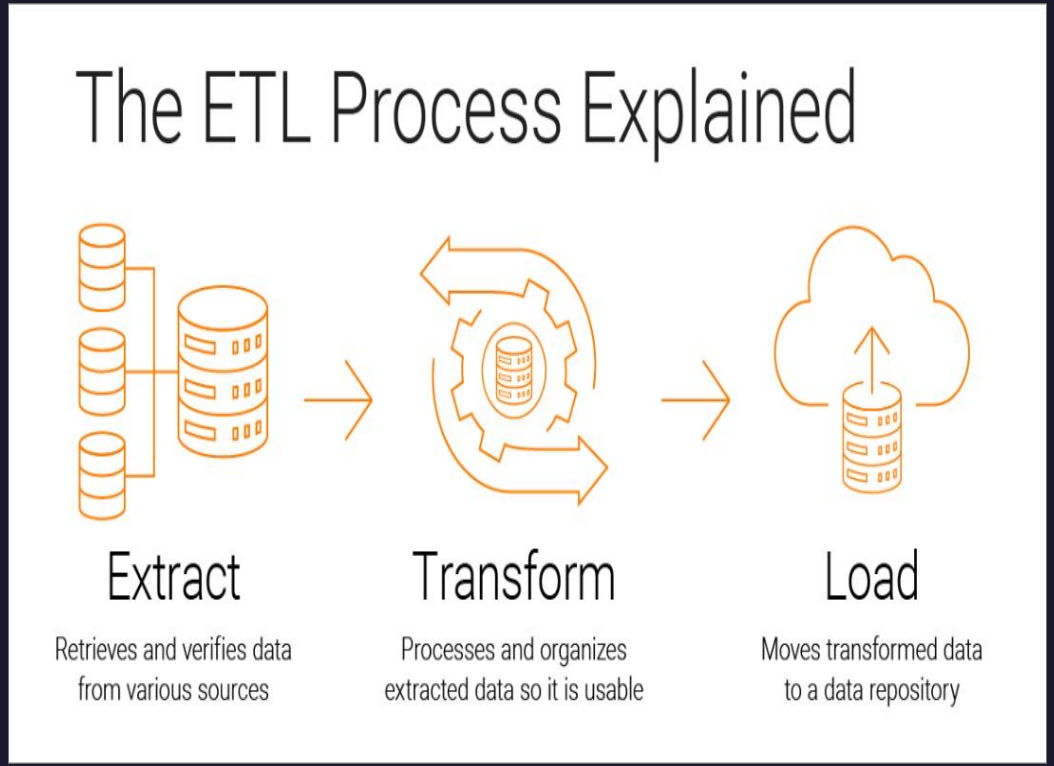
ETL Pipeline Using SQL Database

Libraries used

- pandas
- sqlalchemy
- config

Steps involved

- Extract
- Transform
- Load



ER Diagram


www.quickdatabasediagrams.com
new_country_socioeconomic_transformed
country char(255)
region char(255)
population numeric(12,2)
gdp numeric(10,2)
new_country_olympics_transformed
country char(255)
summer_total integer
winter_total integer
total_participations integer
total_won integer

Prepare Your Database

- Create tables for datasets in the olympics database
  - country_socioeconomic
  - country_olympics
- Primary key
  - country
- Select and inner join the tables

Note: The tables created doesn't have data at this stage.

Extract

- Read the data using pandas library and display the dataframes

	country	summer_participations	summer_gold	summer_silver	summer_b
0	Afghanistan	15	0	0	
1	Algeria	14	5	4	
2	Argentina	25	21	26	
3	Armenia	7	2	8	
4	Australasia	2	3	4	

Transform

- Create a copy of the filtered dataframes from specific columns
- Rename the column headers
- Clean the data by dropping duplicates and NA values if any
- Display the new dataframes

	country	summer_total	winter_total	total_participation	total_won
0	Afghanistan	2	0	15	2
1	Algeria	17	0	17	17
2	Argentina	77	0	45	77
3	Armenia	18	0	15	18
4	Australasia	12	0	2	12

Load

- Connect with the local database
- Inspect tables
- Load the data into the database by using .to_sql method
- Query both tables to confirm if the data has been added

	country	summer_total	winter_total	total_participation	total_won
0	Afghanistan ...	2	0	15	2
1	Algeria ...	17	0	17	17
2	Argentina ...	77	0	45	77
3	Armenia ...	18	0	15	18
4	Australasia ...	12	0	2	12

Join and Preview

- Time to inner join in pandas
- Preview the sql database

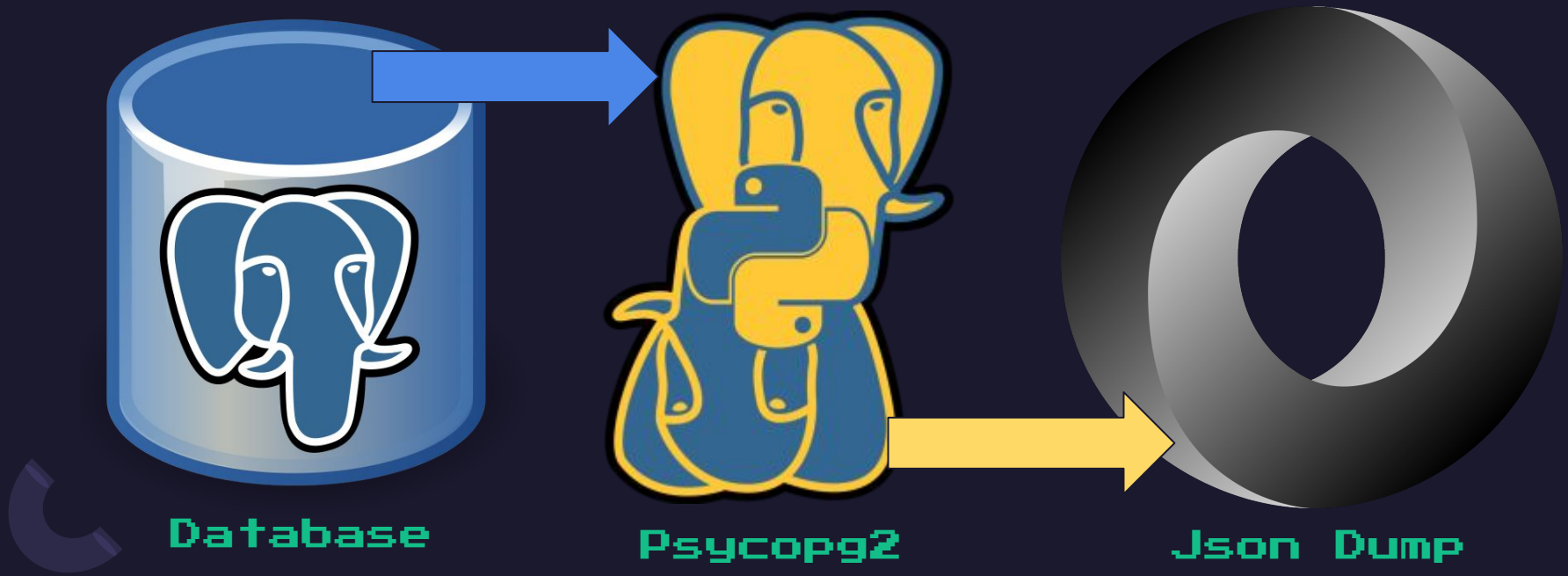
	country	region	population	gdp	summer_total	winter_total	total_participation	total_won
0	Afghanistan ...	SouthernAsia ...	35530	20270	2	0	15	2
1	Algeria ...	NorthernAfrica ...	41318	164779	17	0	17	17
2	Argentina ...	SouthAmerica ...	44271	632343	77	0	45	77
3	Armenia ...	WesternAsia ...	2930	10529	18	0	15	18
4	Australia ...	Oceania ...	24451	1230859	547	19	47	566
...	...	...	...	...	...	...	...	...
122	United States of America ...	NorthernAmerica ...	324460	18036648	2629	330	52	2959

Flask

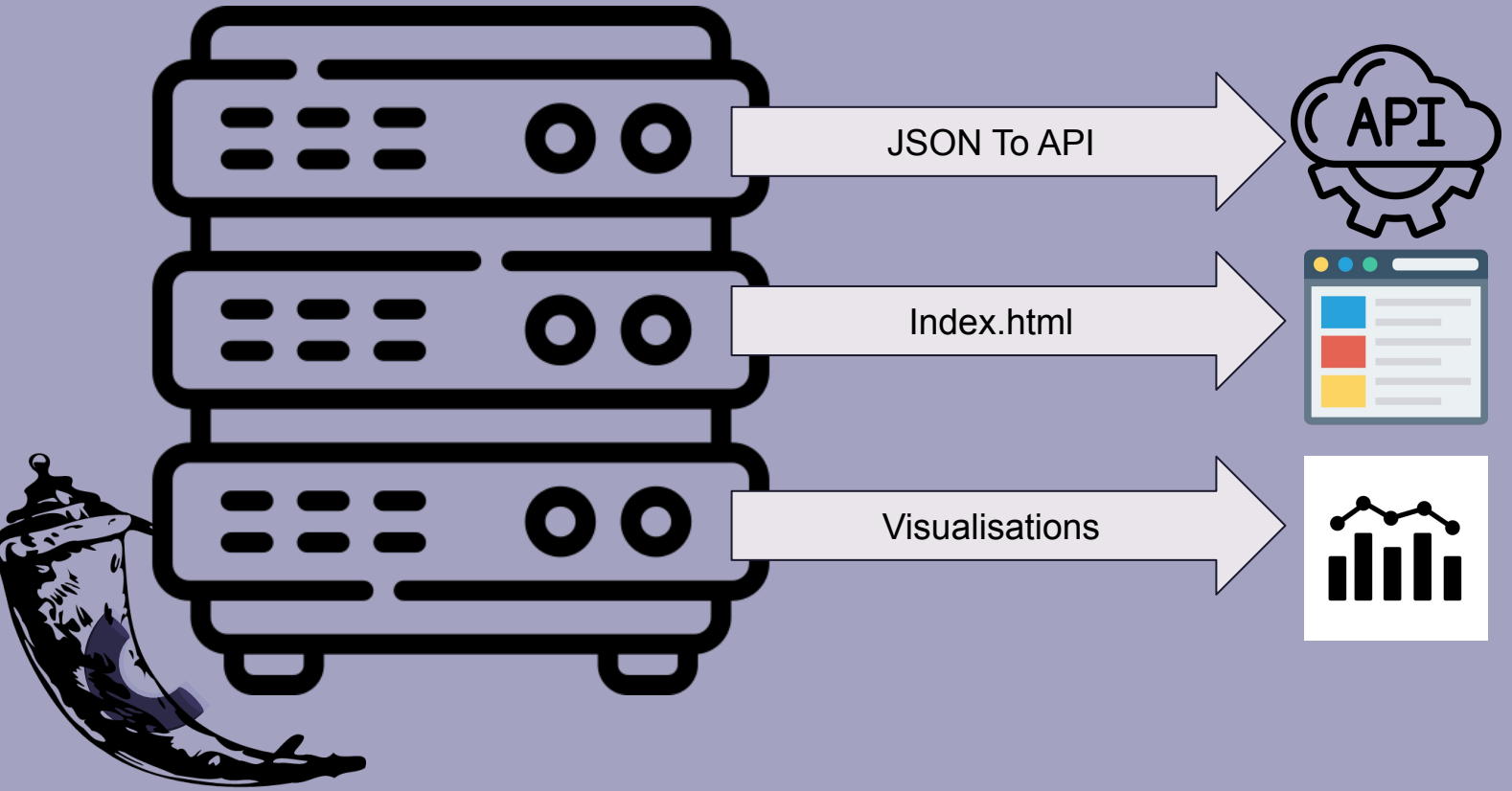
Flask

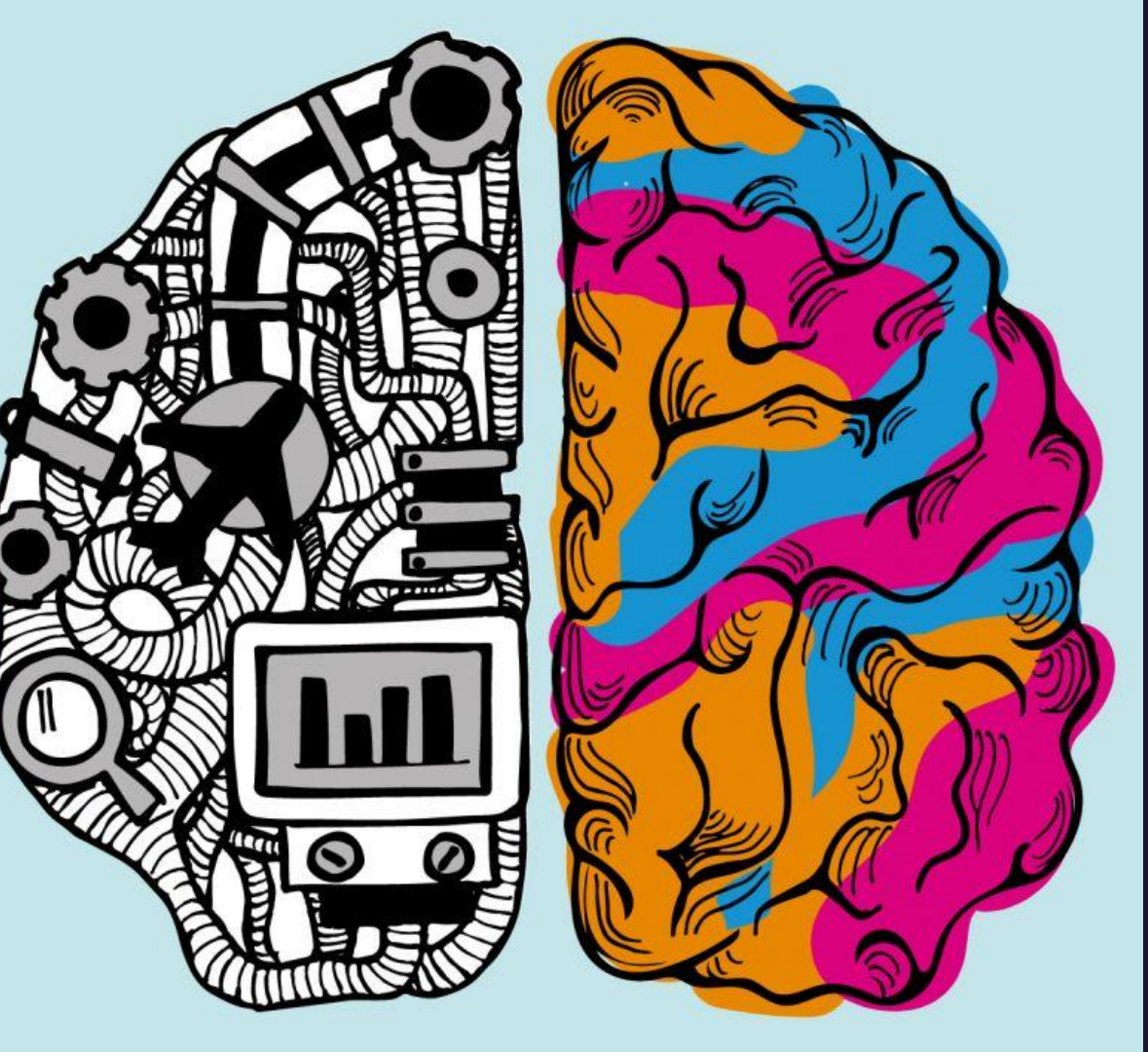
web development,
one drop at a time

Database Retrieval



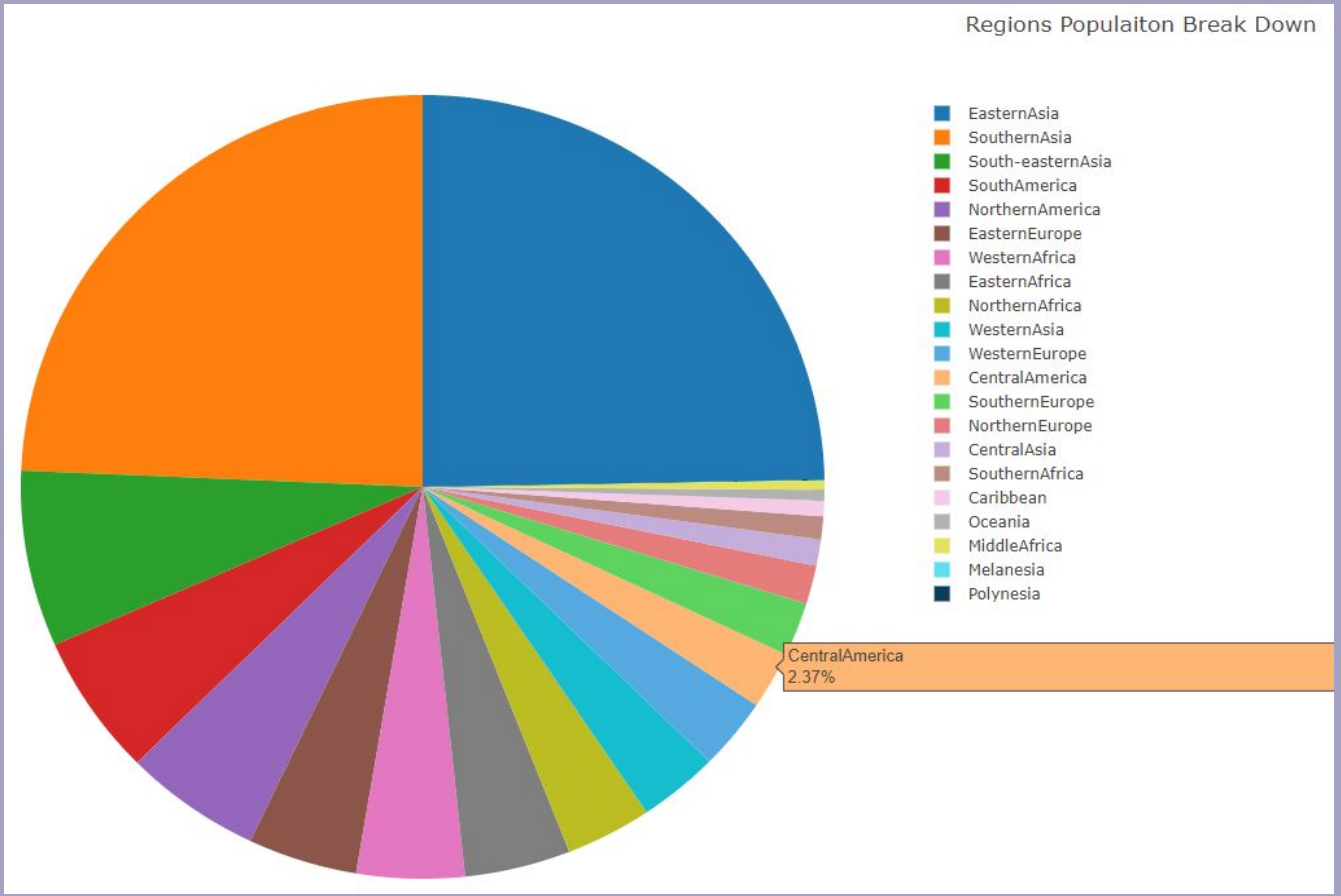
Flask Server Structure



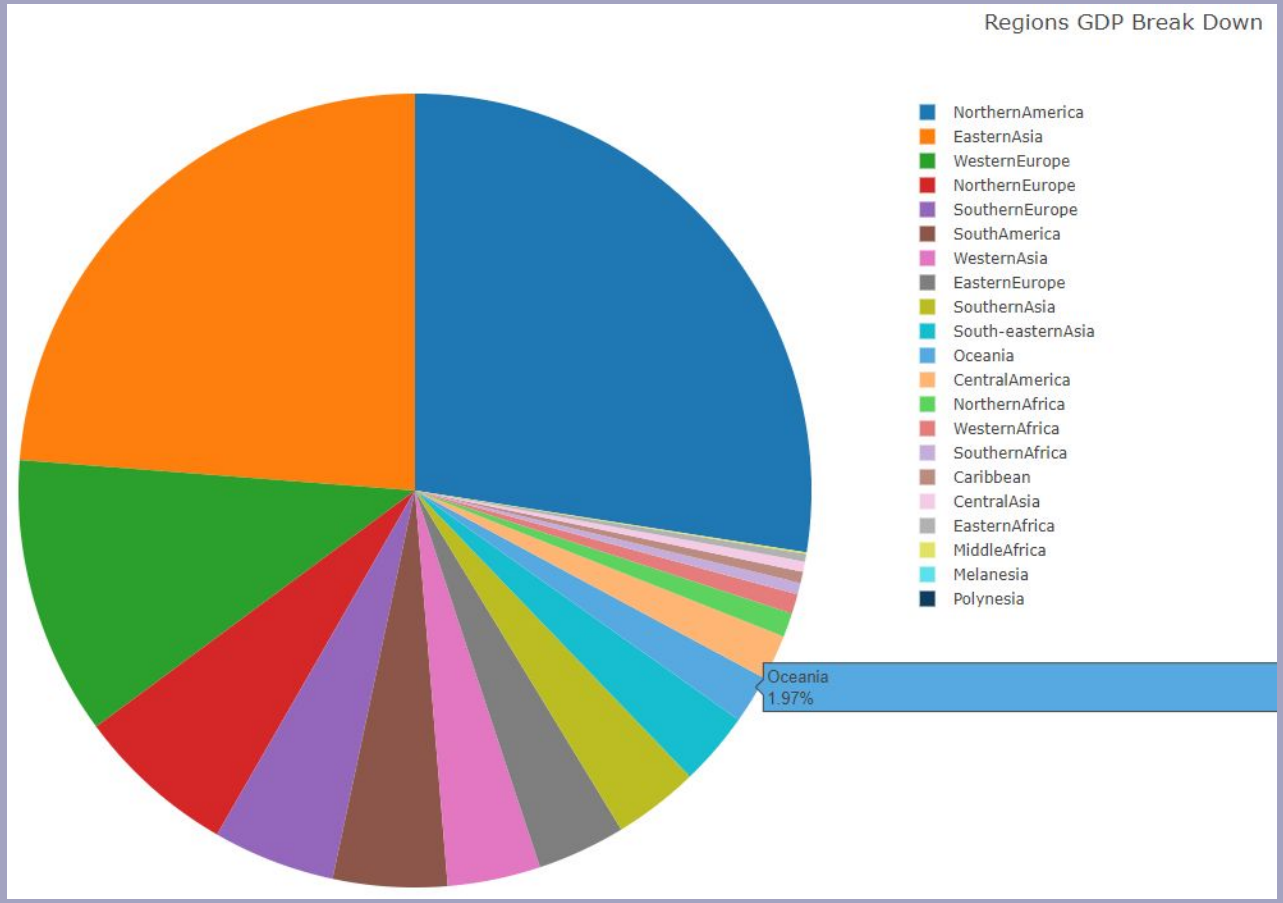


Visualisations

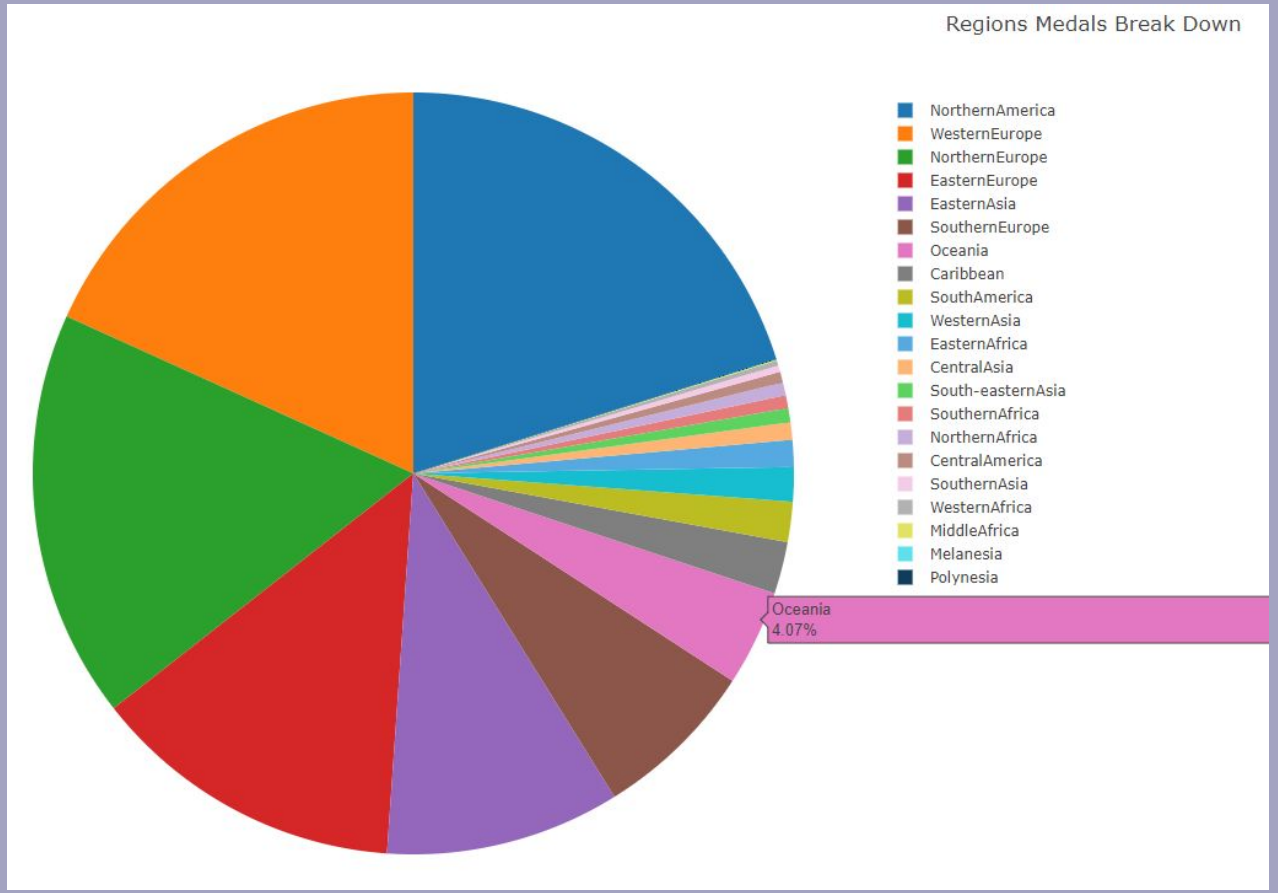
What is the story of the data?



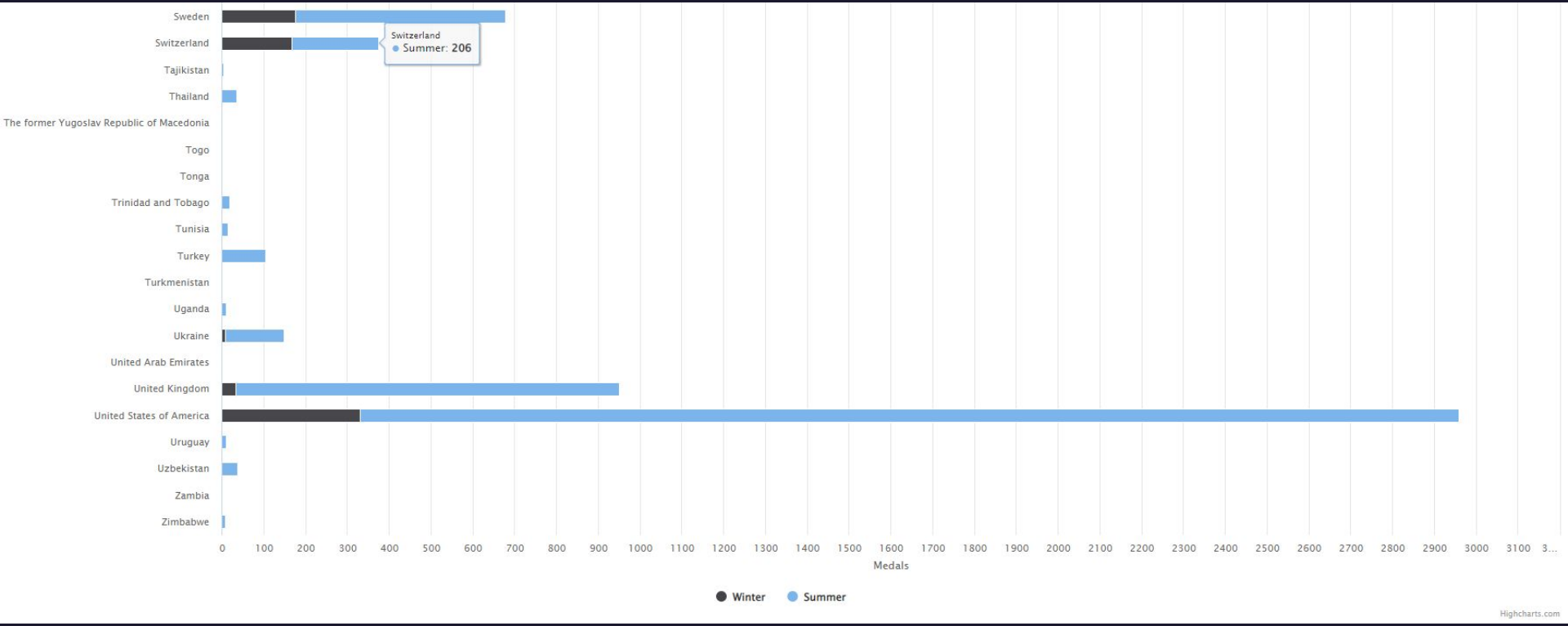
Regional Data - Pie Charts: Population



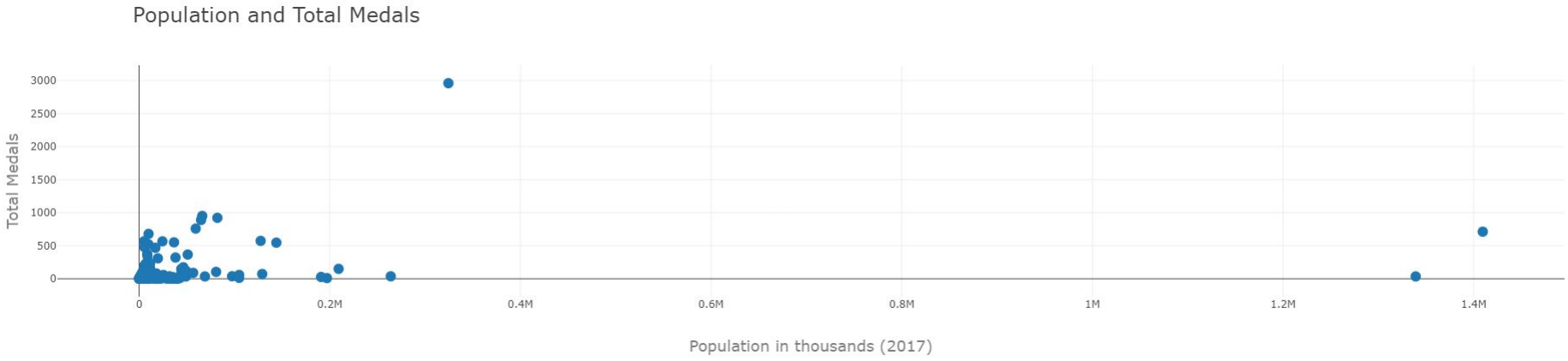
Regional Data - Pie Charts: GDP

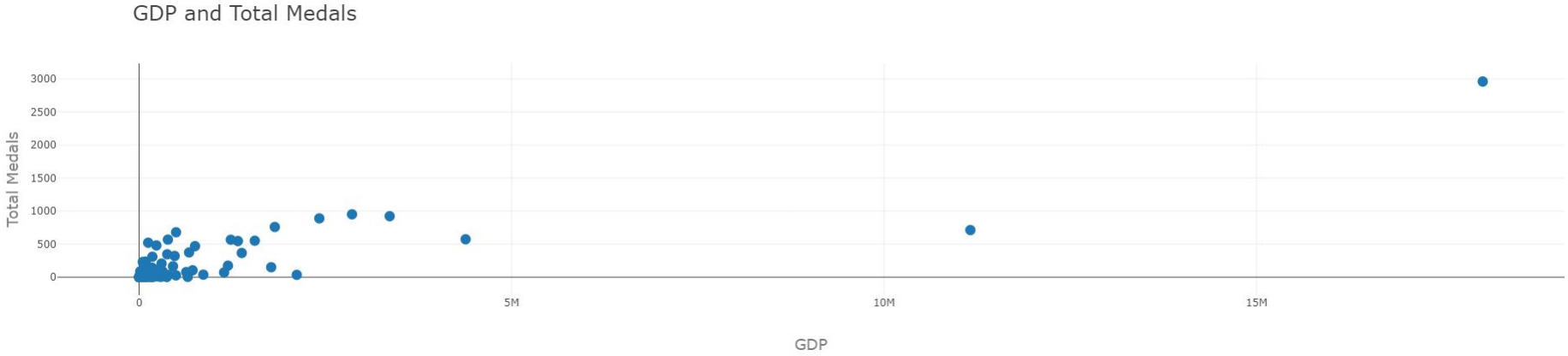


Regional Data - Pie Charts: Medals

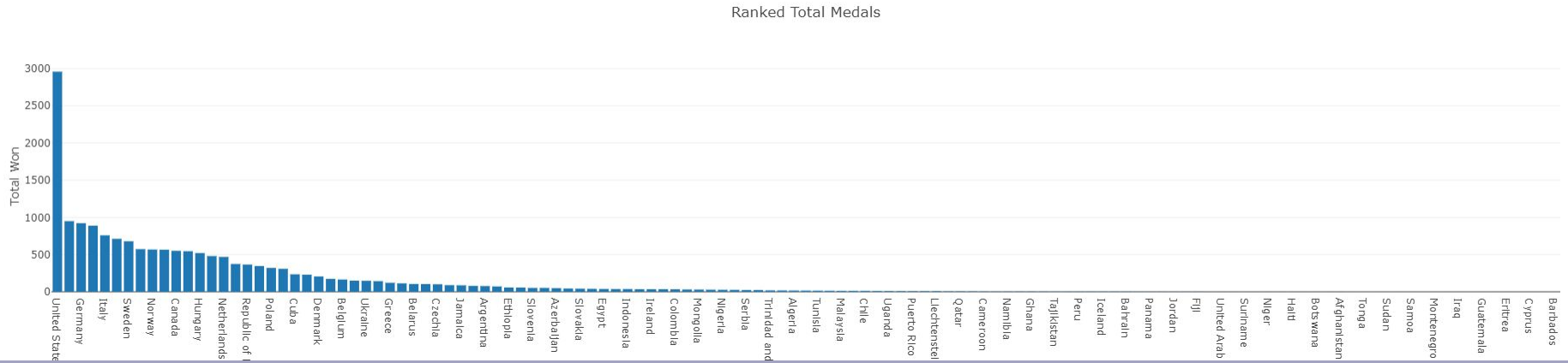


High Chart - Bar Chart: Total Wins

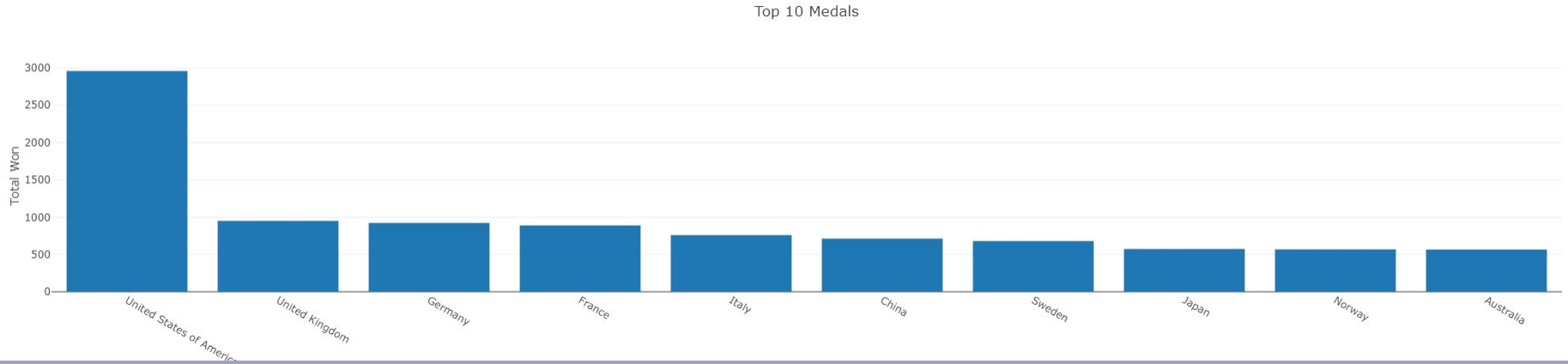




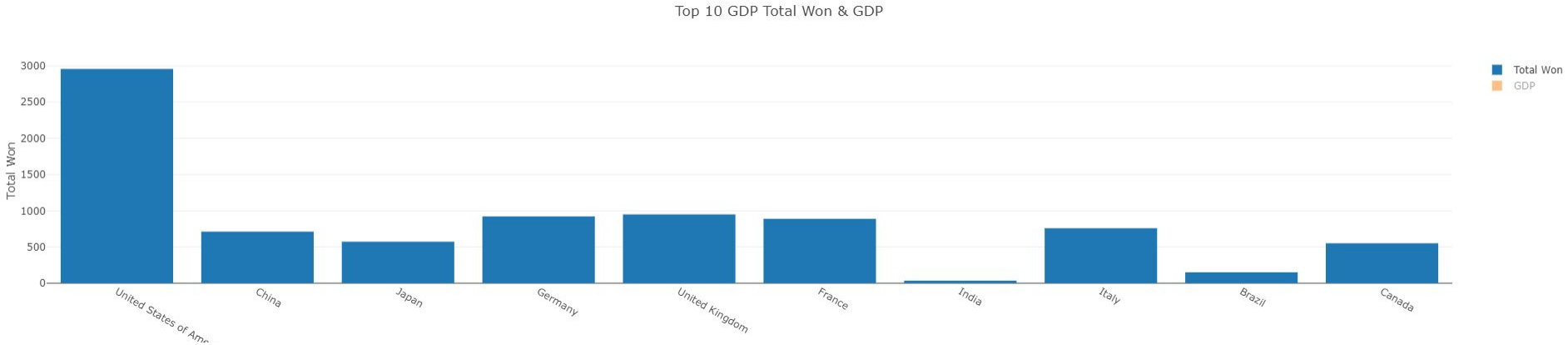
Scatter Plots - Medals With Population & GDP

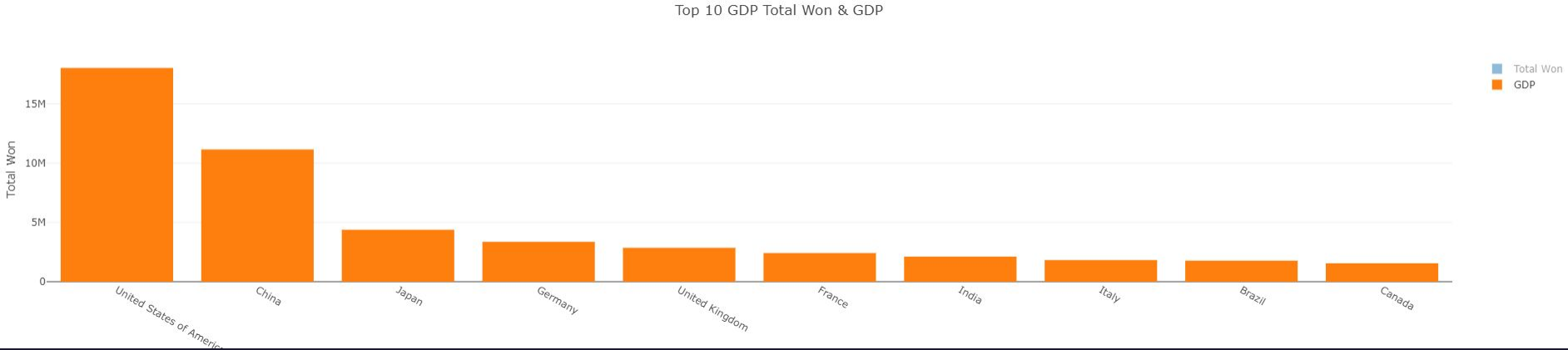


Bar Charts - Ranked

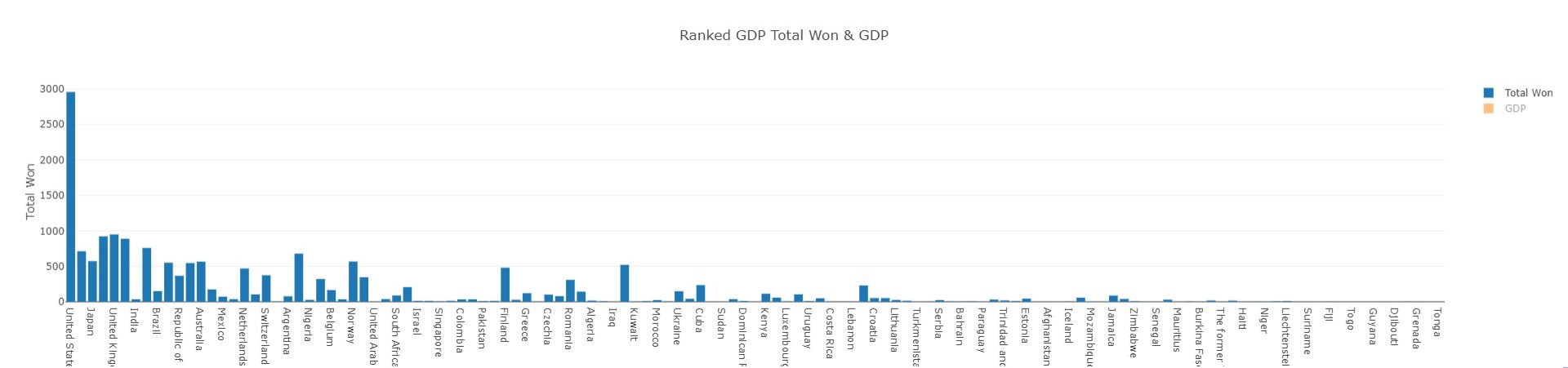


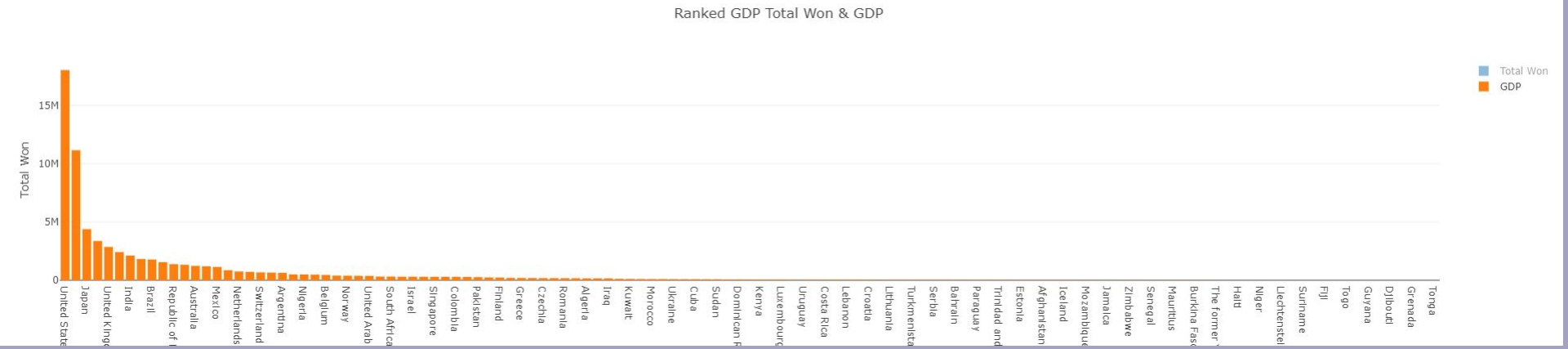
Bar Charts - Top 10



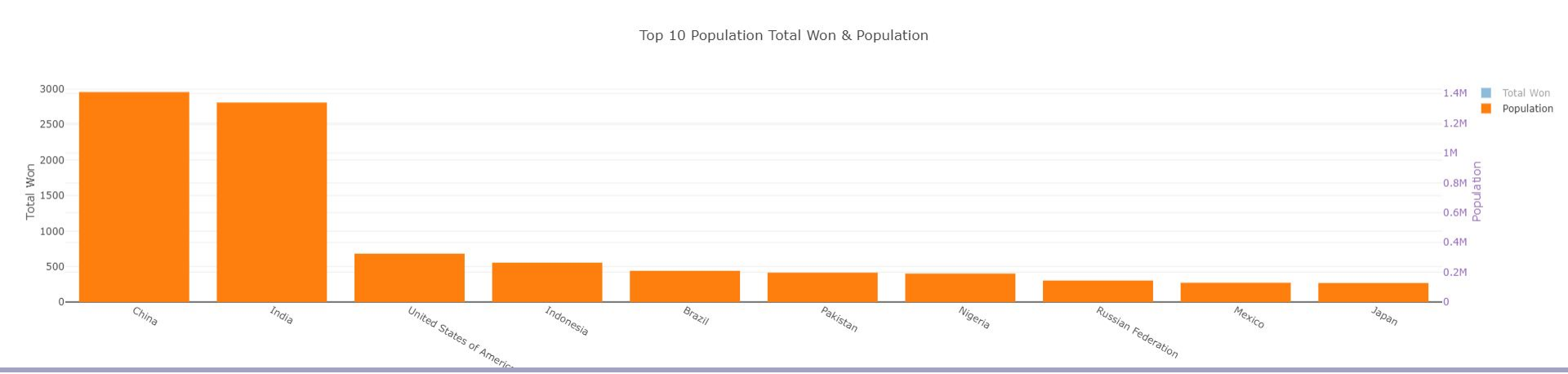


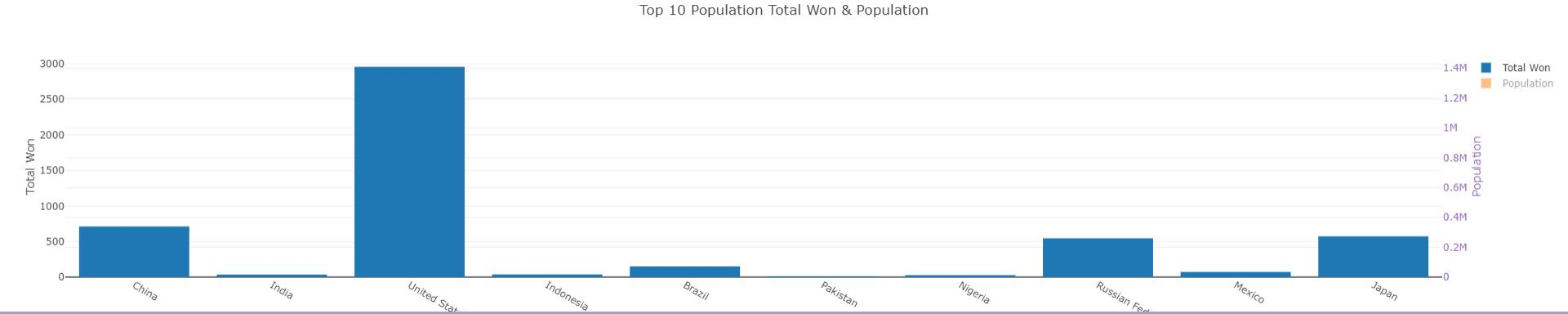
Bar Charts - GDP (2017) Top 10



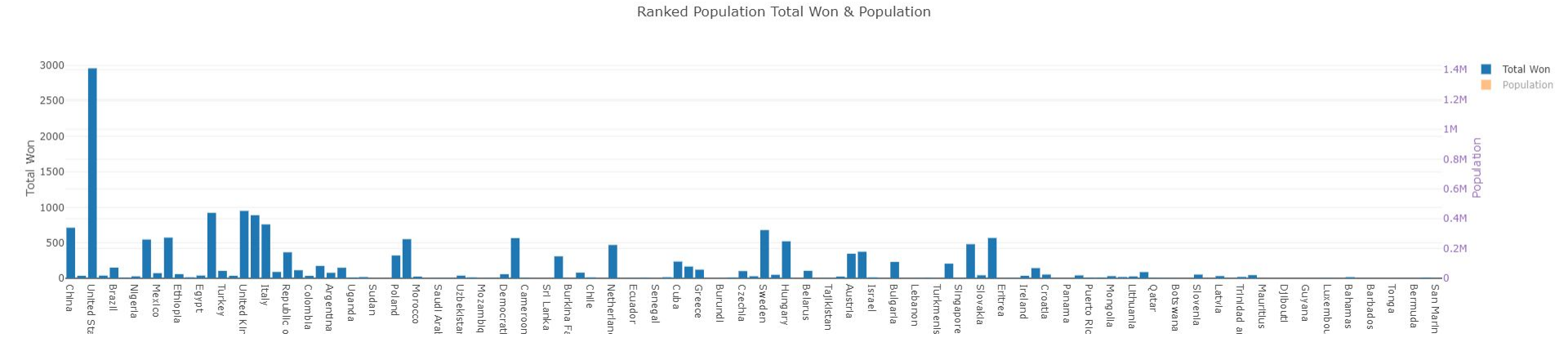


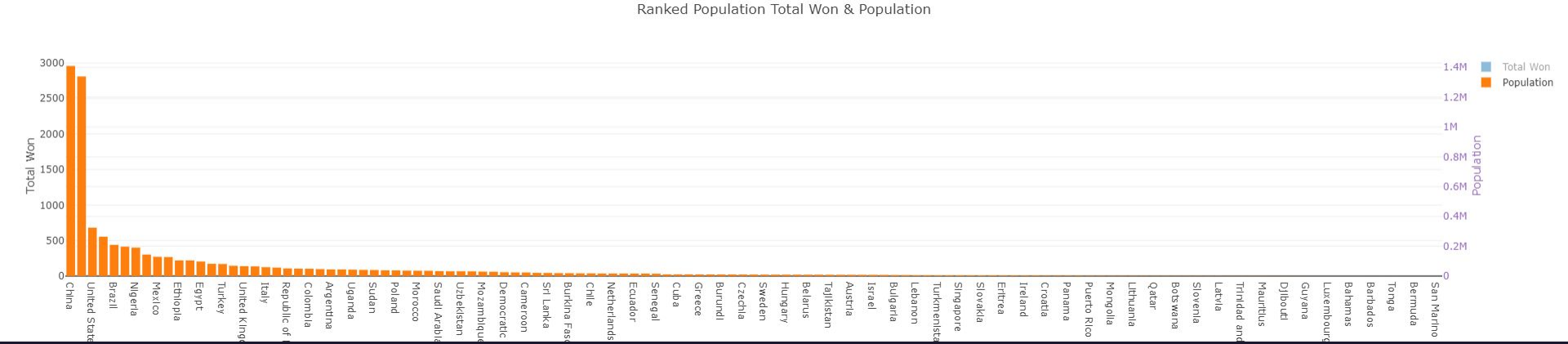
Bar Charts - GDP (2017)



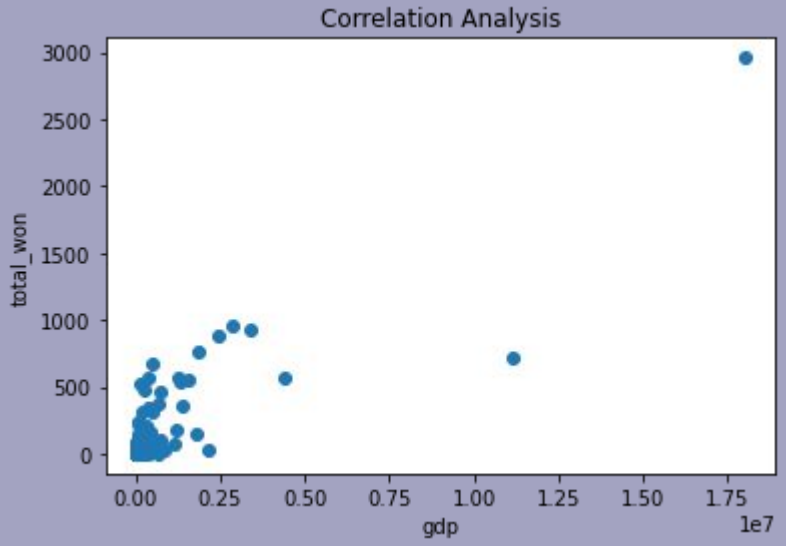


Bar Charts - Population Top 10





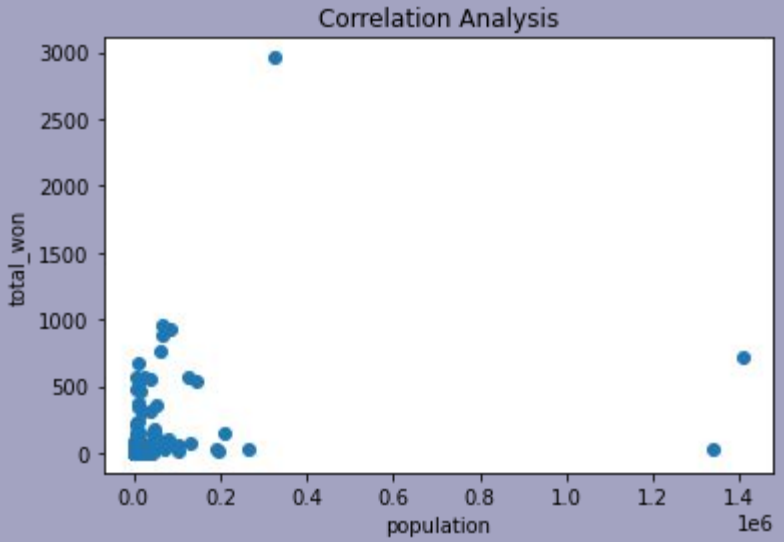
Bar Charts - Population



The correlation between gdp and total medals won is 0.84

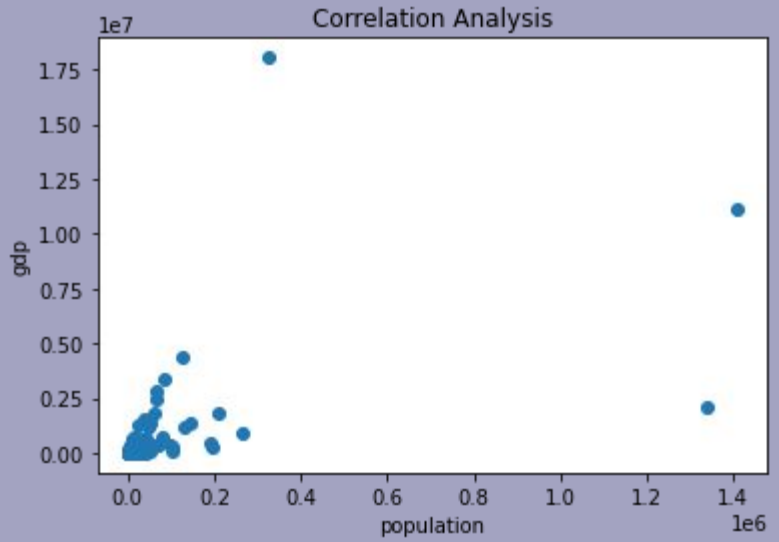
Which is very strong.

Pearson's Correlation	
Absolute Value of r	**Strength of Correlation**
$r < 0.3$	None or very weak
$0.3 \leq r < 0.5$	Weak
$0.5 \leq r < 0.7$	Moderate
$r \geq 0.7$	Strong



The correlation between population and total medals won is 0.23 which is very weak.

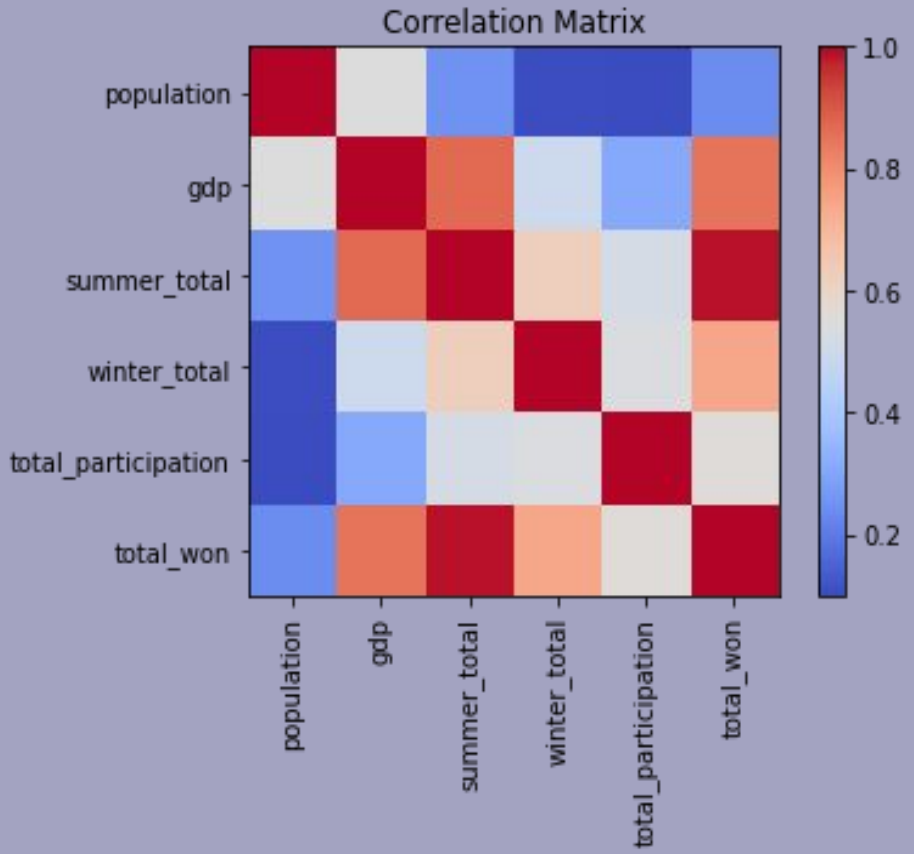
Pearson's Correlation	
Absolute Value of r	**Strength of Correlation**
$r < 0.3$	None or very weak
$0.3 \leq r < 0.5$	Weak
$0.5 \leq r < 0.7$	Moderate
$r \geq 0.7$	Strong



The correlation between population and gdp is 0.54

Which is moderate but not strong.

Pearson's Correlation	
Absolute Value of r	**Strength of Correlation**
$r < 0.3$	None or very weak
$0.3 \leq r < 0.5$	Weak
$0.5 \leq r < 0.7$	Moderate
$r \geq 0.7$	Strong



Pearson's Correlation	
Absolute Value of r	**Strength of Correlation**
r < 0.3	None or very weak
0.3 ≤ r < 0.5	Weak
0.5 ≤ r < 0.7	Moderate
r ≥ 0.7	Strong

Summary

From the analysis I believe:

- GDP has an effect on each country's success in the olympic games.
- Population may or may not have an effect, but much less than GDP.

Resources

1: https://en.wikipedia.org/wiki/All-time_Olympic_Games_medal_table

2: https://www.kaggle.com/datasets/sudalairajkumar/undata-country-profiles

3: https://www.nbcsports.com/bayarea/beijing-2022-winter-olympics/which-countries-have-won-most-olympic-medals

4: https://www.athleteassessments.com/factors-and-obstacles-influencing-olympic-performance/

Thank You

Anam Khalid