Factors influencing Olympics Performance

Peregrin Ryan

Anam Khalid

Prathima

Zitarashe Okot

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!

Agenda

- How does population size affect countries success in winning the Olympic Games?

- What is the relationship between GDP and Olympic medals?

Hypothesis

Our **hypothesis** was to investigate how measurable global factors affect success of countries in the Olympic games.

*We believed that countries with larger **GDP** and **population** would have more success in the Olympic games.*
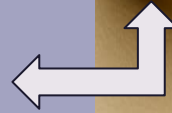
- **What is GDP ?**
  (Gross Domestic Production)

- **How did we measure success in the Olympic games?**

  How did we measure population
  And we had measures medals

  - Population in thousands from 2017
  - winter and summer Olympic games.

ETL Pipeline Using SQL Database

Libraries used

- pandas
- sqlalchemy
- config

Steps involved

- Extract
- Transform
- Load

ER Diagram

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.

country [PK] character (255)	population integer	gdp integer	region character (255)

Extract

- Read the data using pandas library and display the dataframes

	country	summer_participations	summer_gold	summer_silver	summer_
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

Web Scraping - How did we get the Data?

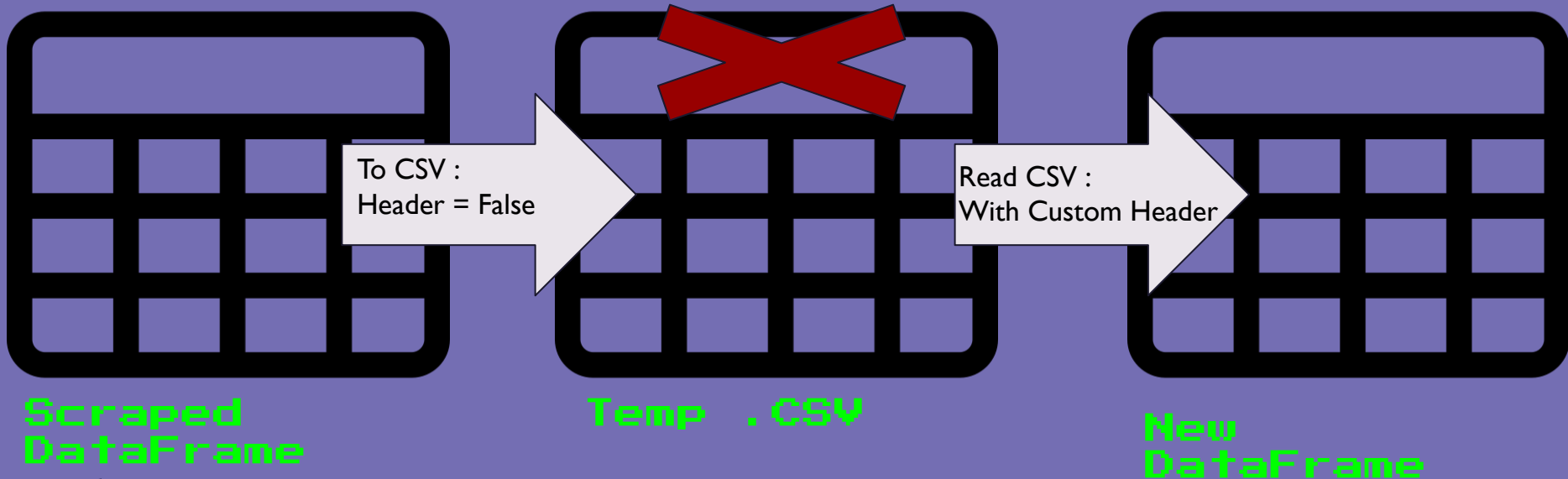# Web Scraping - How did we get the Data?

What was the process?

HTML Websites → Web Scraping → Data
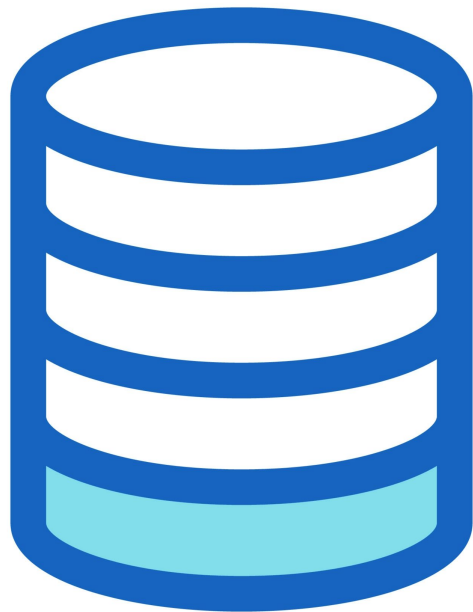
What was the process?
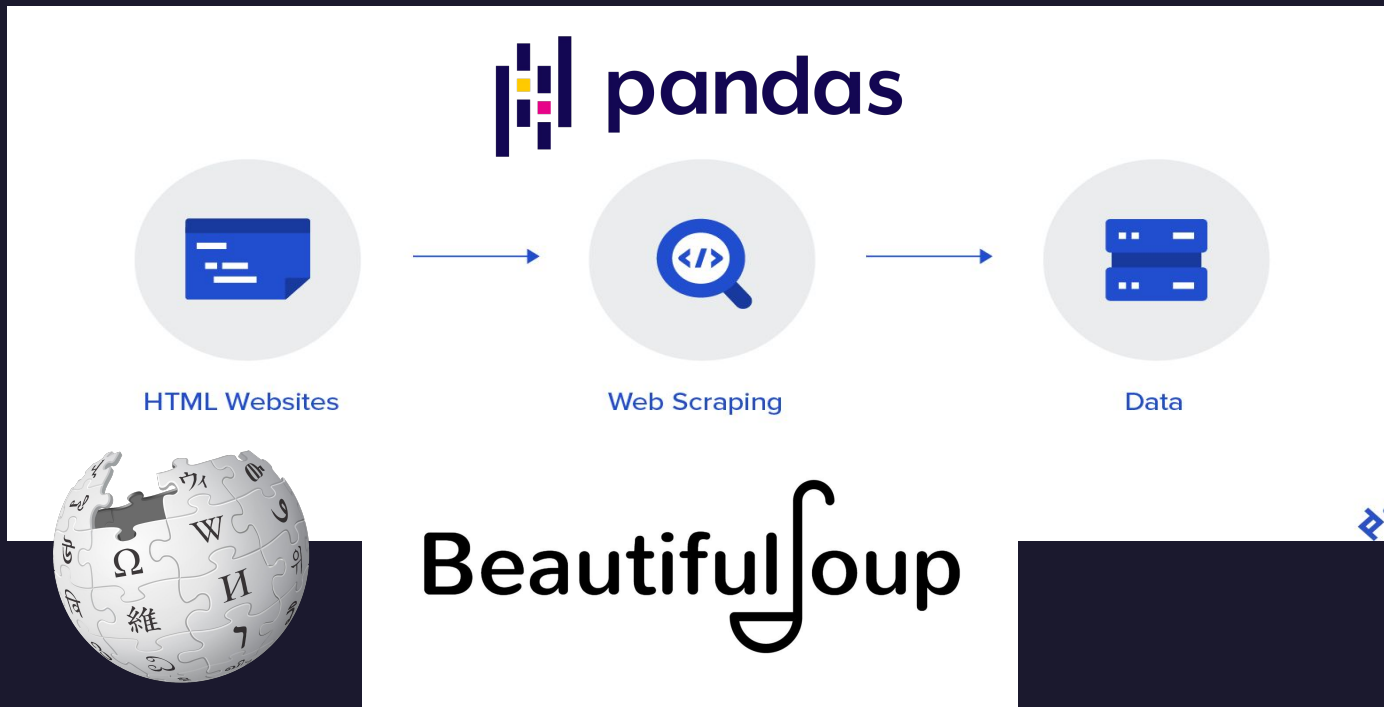
Cleaning the data - What are the problems?

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

The Headers

Scraped
DataFrame

To CSV :
Header = False

OS Module
>>> import os

DELETE Temp.CSV

Read CSV :
With Custom Header

New
DataFrame
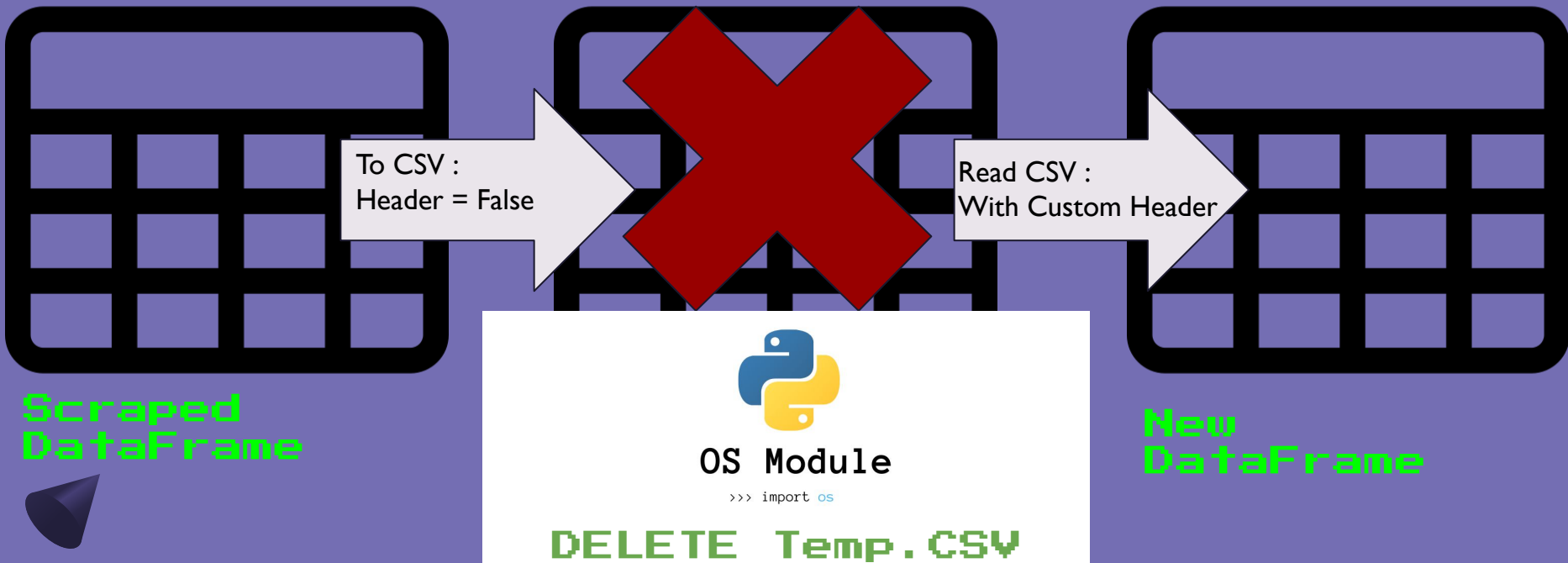
Finishing the cleaning

New
DataFrame

New Headers

country

Afghanistan (AFG)

Algeria (ALG)

Argentina (ARG)

Armenia (ARM)

Australasia (ANZ) [ANZ]

...

Zimbabwe (ZIM) [ZIM]

Delete "(" and everything to the Right
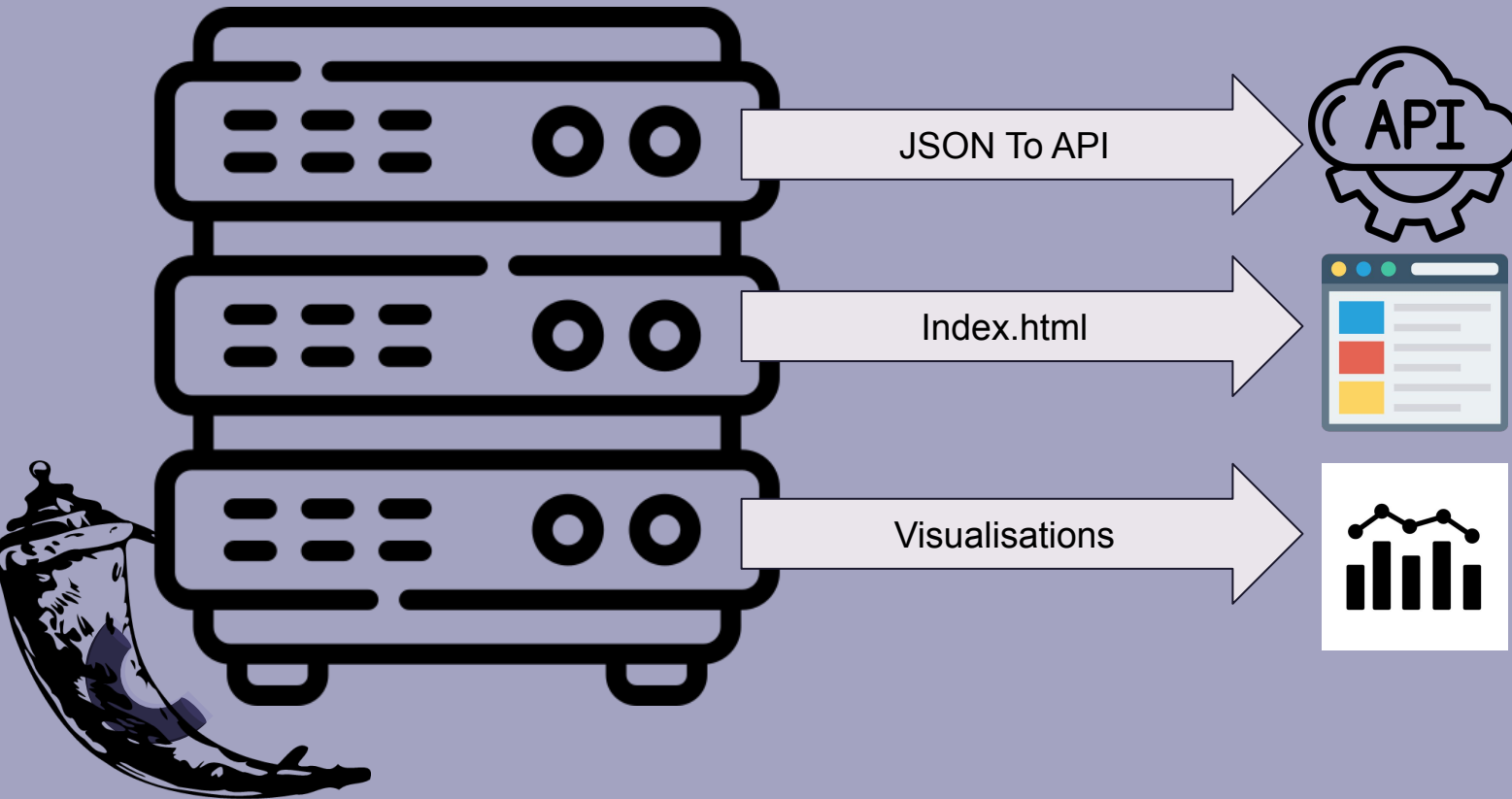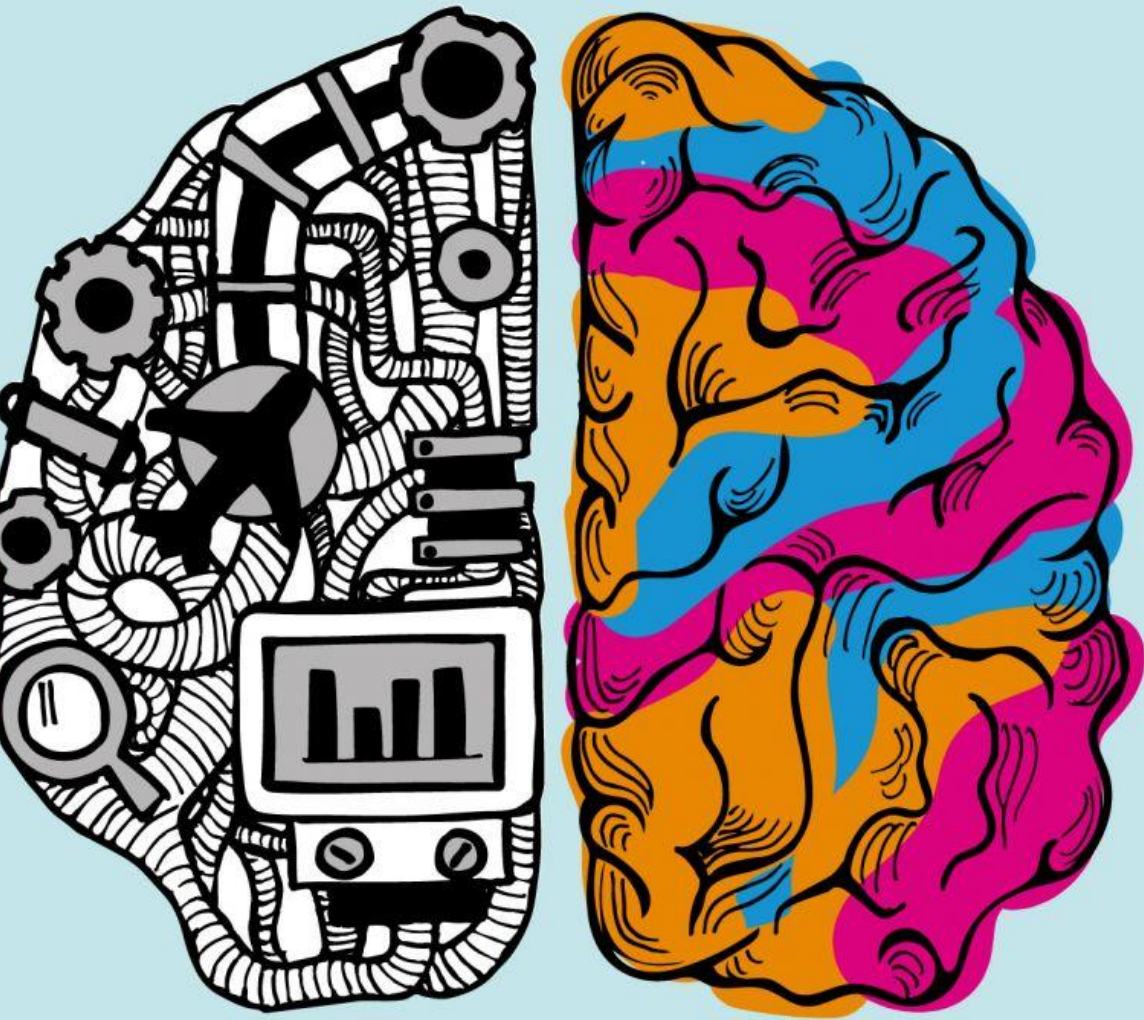
country

Afghanistan

Algeria

Argentina

Flask

Database Retrieval

Flask Server Structure

JSON To API

Index.html

Visualisations

Visualisations

What is the story of the data?

Regions Populaiton Break Down

- EasternAsia
- SouthernAsia
- South-easternAsia
- SouthAmerica
- NorthernAmerica
- EasternEurope
- WesternAfrica
- EasternAfrica
- NorthernAfrica
- WesternAsia
- WesternEurope
- CentralAmerica
- SouthernEurope
- NorthernEurope
- CentralAsia
- SouthernAfrica
- Caribbean
- Oceania
- MiddleAfrica
- Melanesia
- Polynesia

CentralAmerica
2.37%
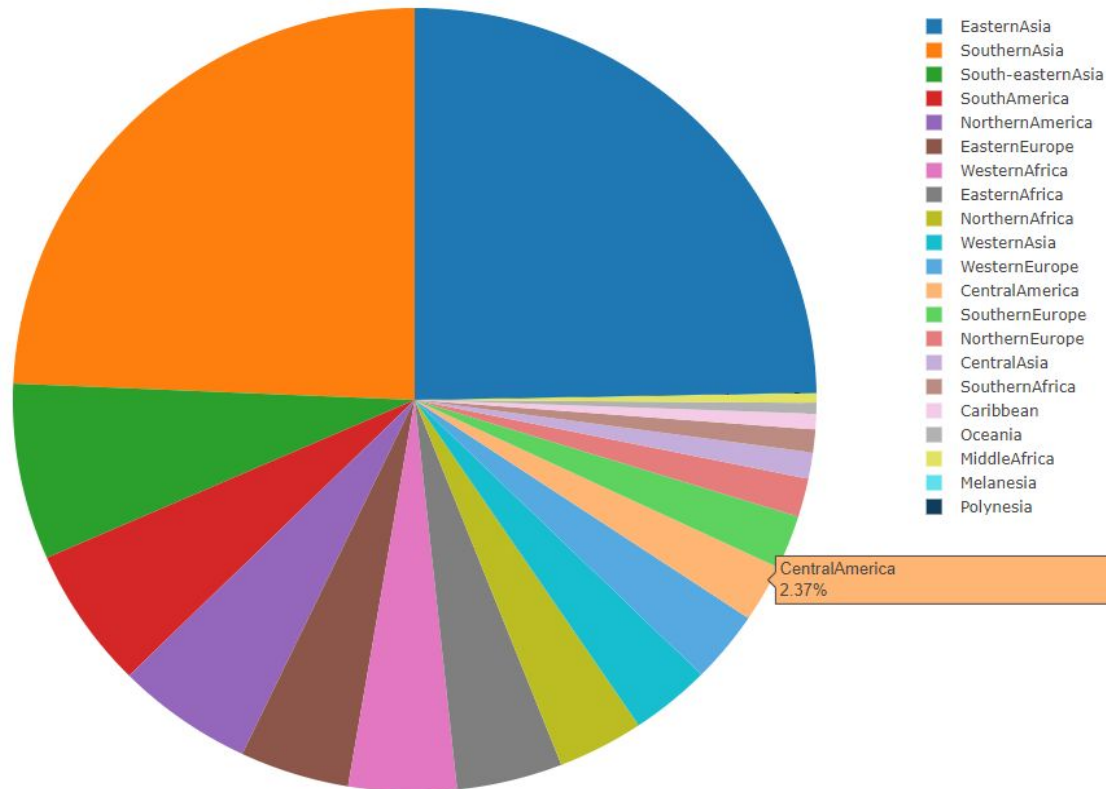
Regional Data - Pie Charts: Population

Regions GDP Break Down

- NorthernAmerica
- EasternAsia
- WesternEurope
- NorthernEurope
- SouthernEurope
- SouthAmerica
- WesternAsia
- EasternEurope
- SouthernAsia
- South-easternAsia
- Oceania
- CentralAmerica
- NorthernAfrica
- WesternAfrica
- SouthernAfrica
- Caribbean
- CentralAsia
- EasternAfrica
- MiddleAfrica
- Melanesia
- Polynesia

Oceania
1.97%
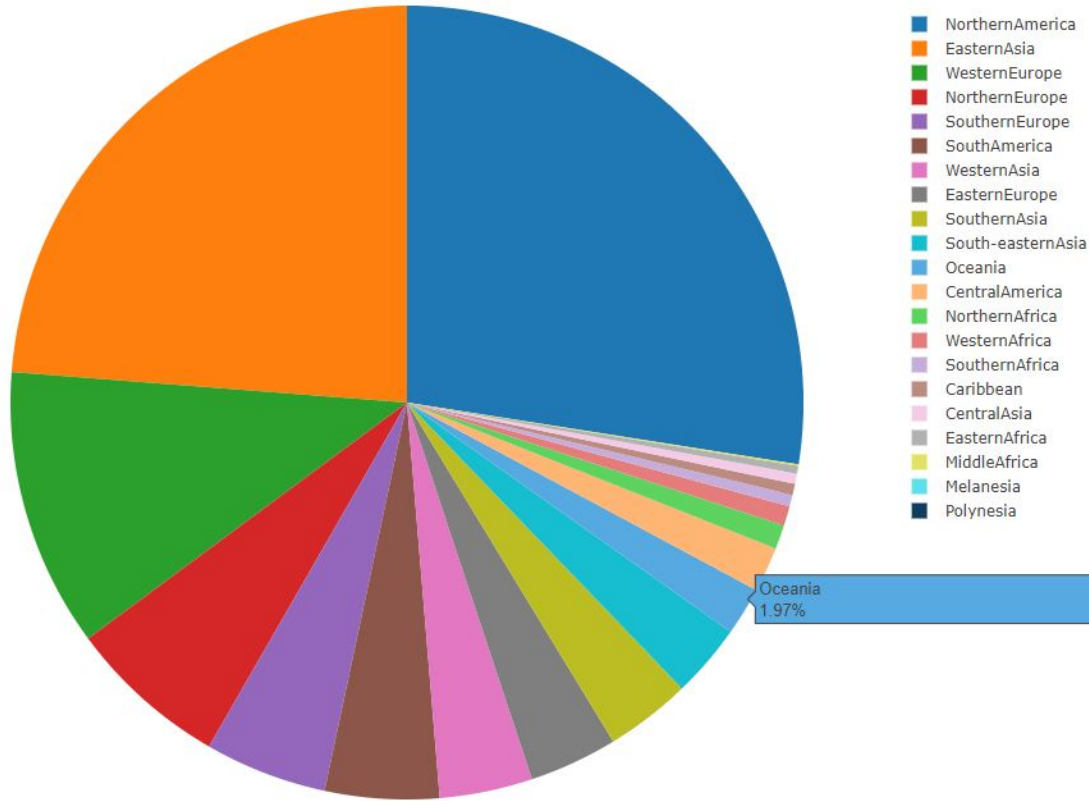
Regional Data - Pie Charts: GDP
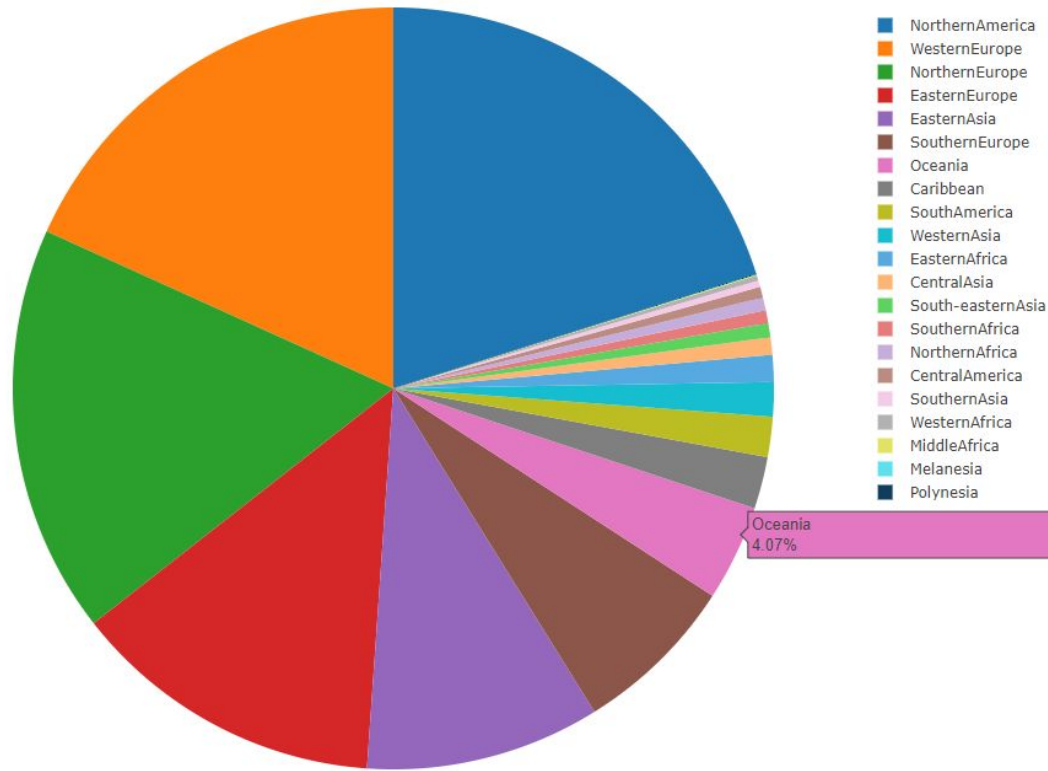
Regional Data - Pie Charts: Medals
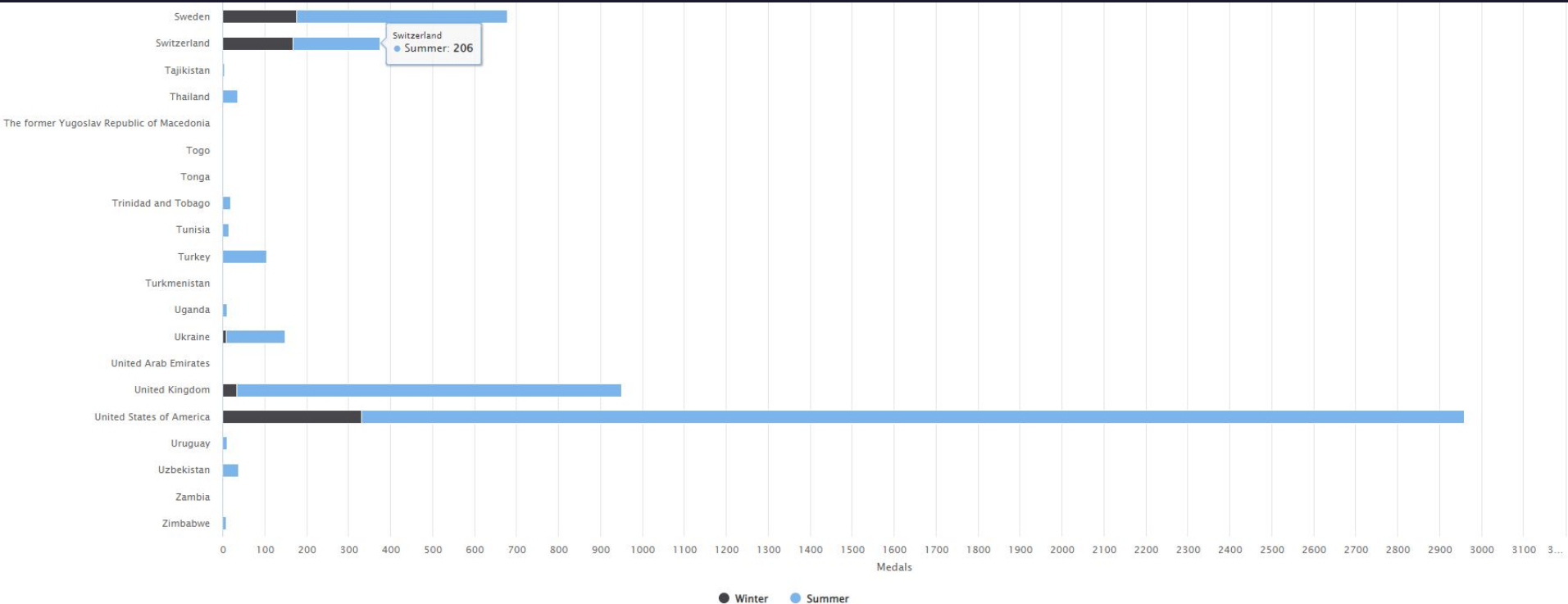
High Chart - Bar Chart: Total Wins

Population and Total Medals
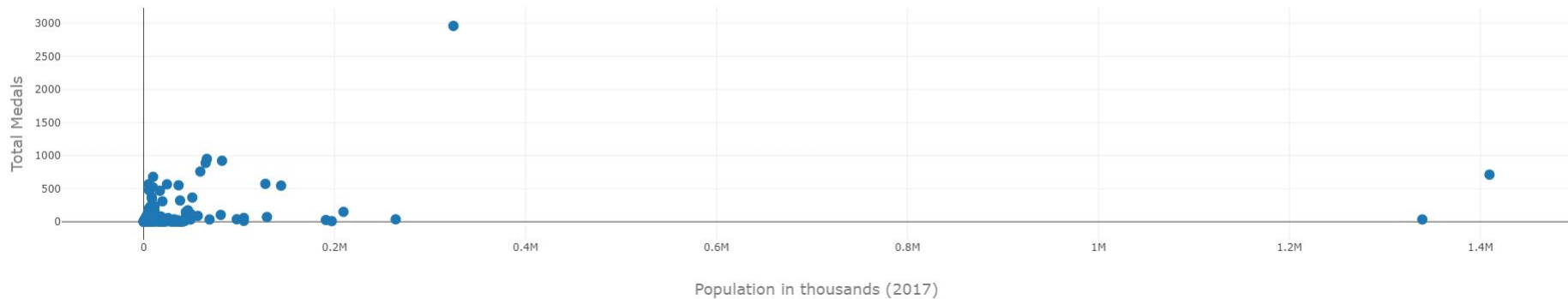
GDP and Total Medals
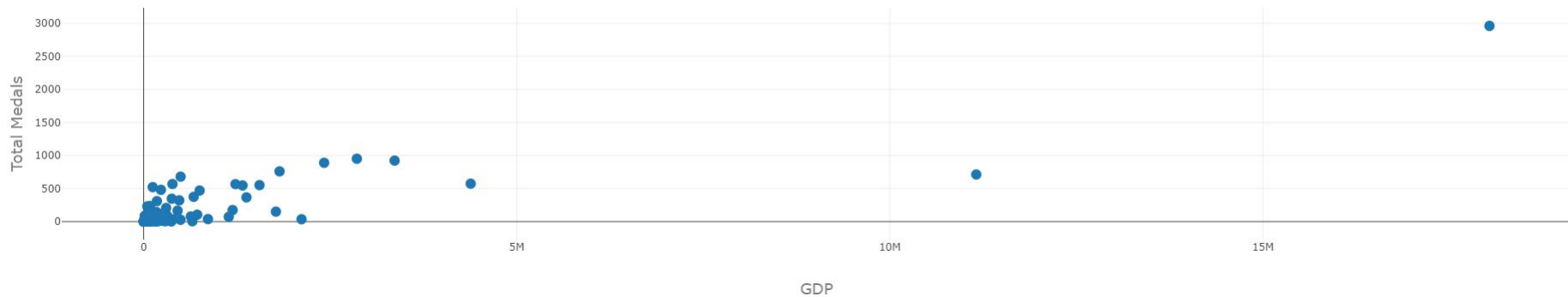
Scatter Plots - Medals With Population & GDP
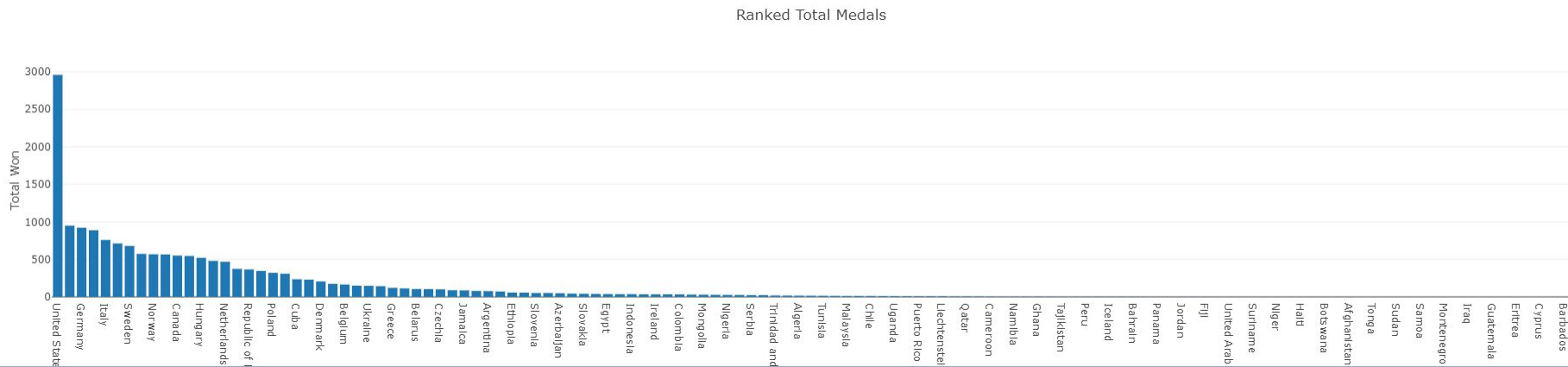
Bar Charts - Ranked

Top 10 Medals
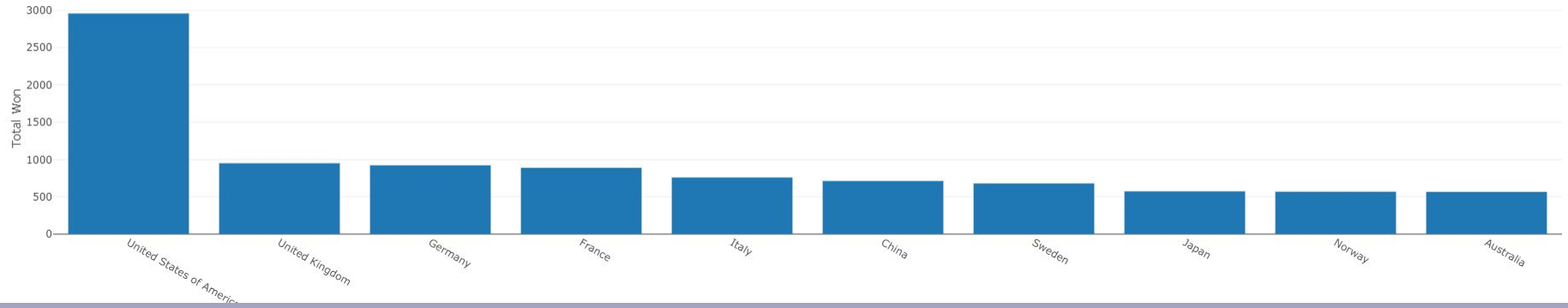
Bar Charts - Top 10
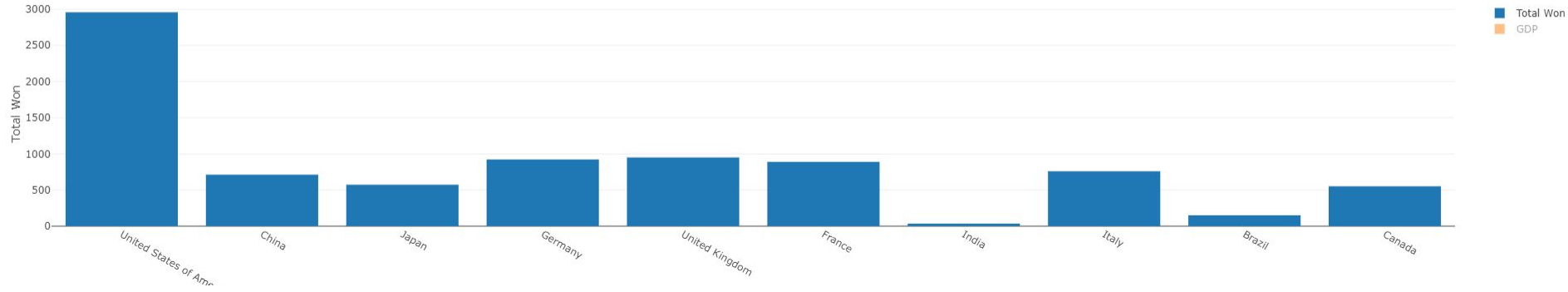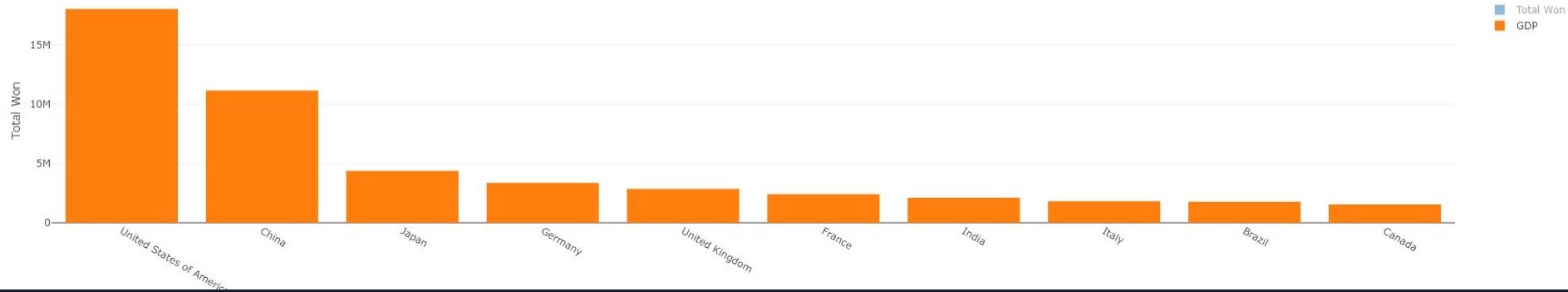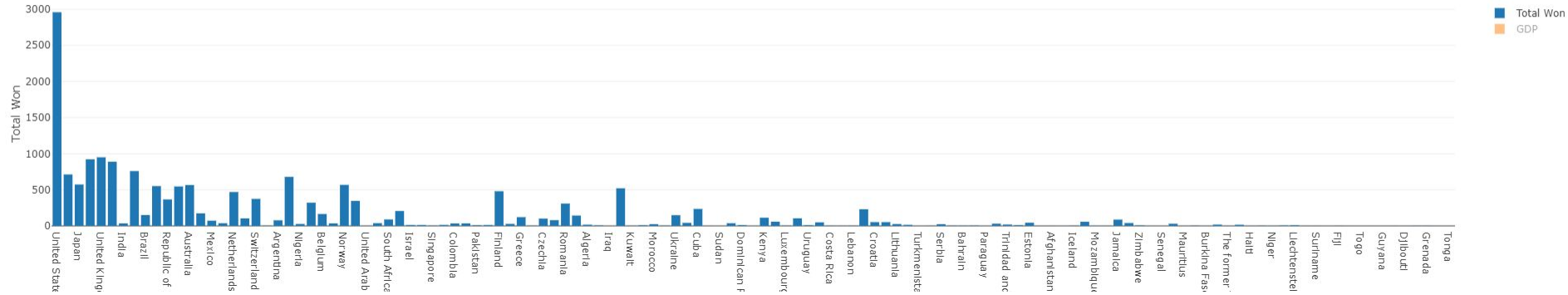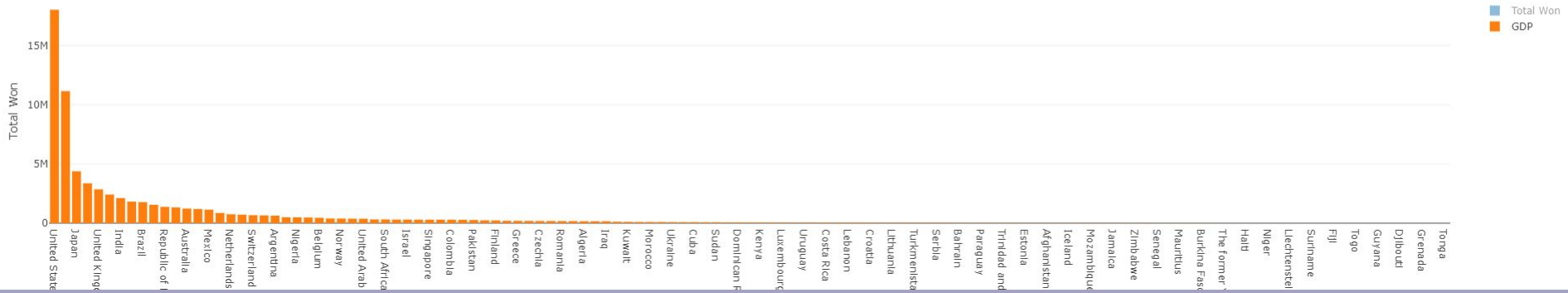
Bar Charts - GDP (2017) Top 10

Bar Charts - GDP (2017)
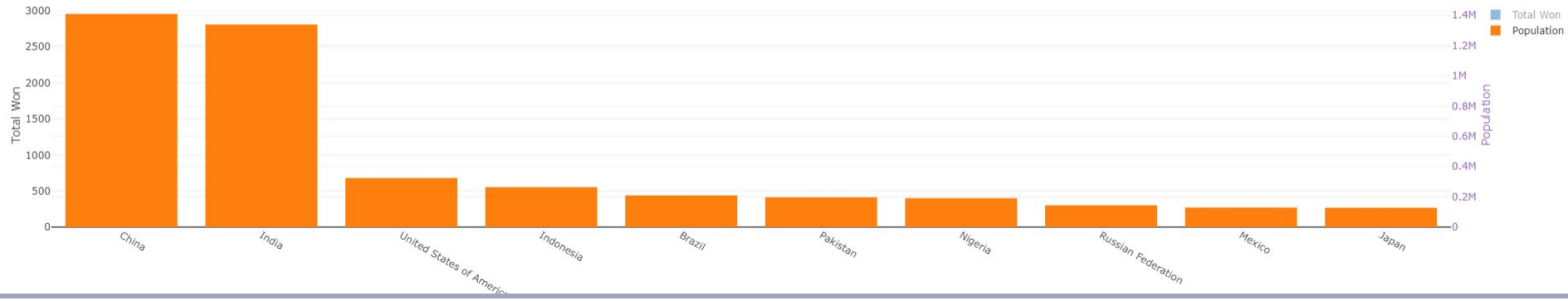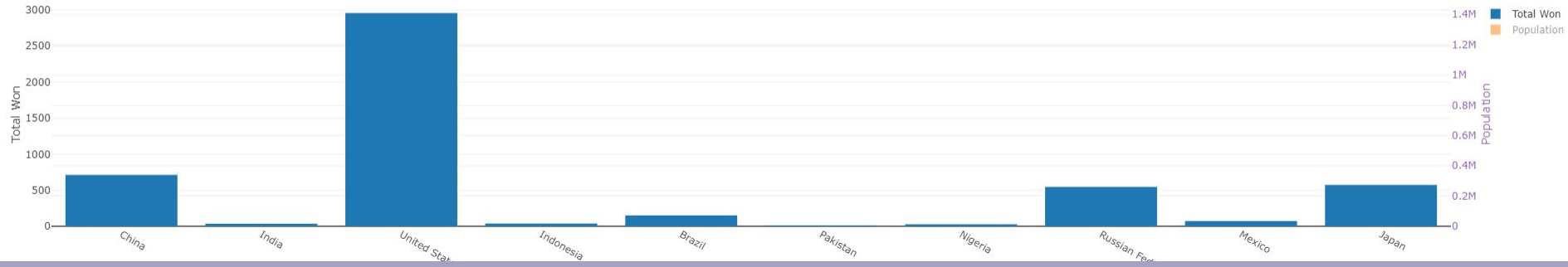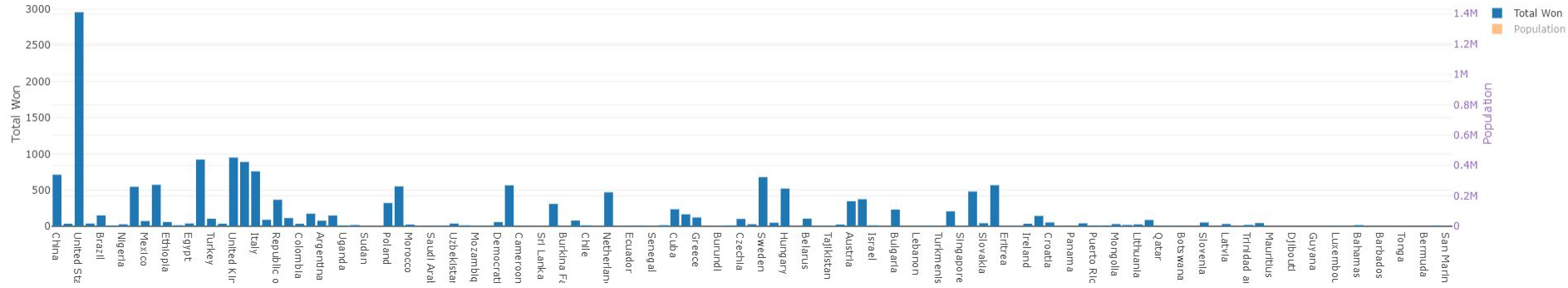
Bar Charts - Population Top 10
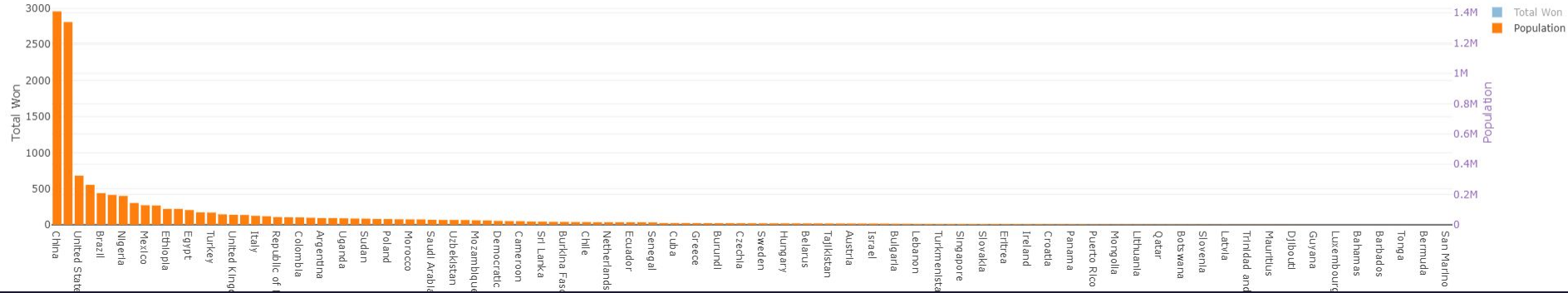
Bar Charts - Population

Summary

From our analysis we believe:

- GDP has an effect on each countries success in the olympic games.
- Population may have an effect, but much less than GDP.

Project 3

Thank You

Peregrin Ryan

Anam Khalid

Prathima

Zitarashe Okot

Resources

[What is GDP and how is it measured? - BBC News](#)

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

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