Olympic Games Data Analysis

Problem statement and why the problem you are trying to solve interesting/real world implications and/or applications:

We propose to analyze data of the Olympic Games in history and visualize the relationship of competition results with athlete genders, body types and ages. We will also analyze some basic trends such as the number of athletes and the number of medals together with the history background behind them. Based on the analysis, we recommend body types and suggest the most effective regions for athletes to train in to achieve excellence in specific sports.

What kind of data set you plan on scraping/collecting:

Kaggle Datasets:

https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results

https://www.kaggle.com/the-guardian/olympic-games

These two datasets contain 4 csv files with body types and genders of athletes, medal results and so on in Olympic Games.

Systematic plan/approach mentioning division of labor and effort and time table

1. Put data from the different sources into Pandas Dataframe with the following data separated into columns: Name, Sex, Age, Height, Weight, Team, Year, Season, City, Sport, Event, Medal.
2. Use 2 sided unpaired t test to compare the above mentioned variables, using country of origin as the grouping factor. This is a very simple statistical analysis method that only takes in 2 variables. If however we wanted to look at the correlation between weight, number of medals and country, we will have to use multivariate statistical analysis methods. For this example we might need to use Multivariate Analysis of Variance (MANOVA). The main challenge in this project will be using the correct multivariate statistical analysis method for the right sets of variables.
3. We plan on making a gif of medal distribution around world in a period of years using Matplotlib and Basemap. To fully complete such animation show, we need learn how to use functions in matplotlib.animation and accommodate it with basemap package.

Project steps	Time for completion	Person(s) in charge (among the group of 12)
1. Extract and clean the data.	1 week	Rui Cao
2. Perform statistical analysis.	2 weeks	Nikhil Goel

3. Data Visualization	1 week	Ke Han