Title: Tokyo Olympics 2020: Unveiling the Data Behind the Games

1. Project Summary

OlympicData is a comprehensive application that allows users to search and visualize data from the 2021 Tokyo Olympics. By using data about athletes, coaches, gender, teams and medals, the application offers three different pages: Athletes, Teams, Medals where users can search and visualize data related to the Olympics based on the page they go on. Moreover, there is a world map visualization that is offered to the users where they can hover over each country that participated in the Olympics to see the number of medals won, participants, and coaches. This allows users to get a detailed view of the world and see who participated while also making it easy to use. Overall, our goal is to create a web application that is intuitive and also satisfies any queries related to the Tokyo Olympics.

2. Detailed Description

1. Dataset: The data stored in the database is from the Kaggle Dataset: 2021 Olympics in Tokyo. The data is from the previous Olympics two years ago in Tokyo. There are five key Excel sheets that are stored in the dataset: Athletes, Coaches, EntriesGender, Medals, and Teams. Here are the respective columns in each sheet which will be transferred to a table per sheet:

Athletes: Name, NOC, Discipline
Coaches: Name, NOC, Discipline, Event
EntriesGender: Discipline, Female, Male, Total
Medals: Rank, NOC, Gold, Silver, Bronze
Teams: Name, Discipline, NOC, Event

We would store all this information in different tables and use all this information in our application.

2. Functionality: The basic functions of our application are to be able to search/query information related to each table and also explore visualizations related to each table. The three pages offered to the users would be Athletes and Coaches, Teams and Medals, each of which the user would be able to see visualizations and also search through the tables to see their query. A simple feature is to search through each table using a search bar and the data is sorted based on the query. A complex feature that we added is to create visualizations such as histograms for each page. Examples would be number of participants per country, number of females/male/total participants in each event, etc. We want to create visualizations that the users can interact with per page.

3. Creative component that we thought would be interesting is to add a world map that users can interact with in order to see the number of medals, participants, and coaches per country. This would help users be able to sort a large amount of data quickly by just hovering over a country and seeing how the country did generally. We are planning to do this by using an interactive map library such as Visme and JavaScript to create the map and publish it onto our front end.

3. **Usefulness**: The current official olympics website provides a comprehensive overview of various olympic games over the years and an ability to look up any information regarding the olympic games. However, it doesn't provide
   1. Data visualizations/ Descriptive statistics: The purpose of having some data visualization is to easily comprehend complex or vast data which if a user has to search and draw conclusions would be a tedious task. Descriptive analytics and visualization improves interpretability of the data. It also provides an easy means to remember the data.
   2. Inferential statistics: Inferential statistics would assist countries/ coaches/ athletes to invest time and effort in the right sport to improve their ranking in the future events.

   For instance, having a bar chart representation for the total number of medals in each sport in a country would help the country understand the disciplines that they are strong and weak at. They could invest more time in the weak disciplines to improve it.

4. **Realness**

   The dataset we will be using for this project is the 2021 Olympics in Tokyo from kaggle. The link to the dataset on kaggle:.

   The dataset consists of 5 tables; all in the xlsx (XML format). The 5 tables and their schema is as follows:

   1. Athletes
      a. (Name, NOC(National Olympics Committee), Discipline)
         i. Name: Name of the athlete
         ii. NOC: The country the athlete belongs to
         iii. Discipline: The sport that the athlete participated in
      b. The table has 11062 athletes.

   2. Coaches
      a. (Name, NOC(National Olympics Committee), Discipline, Event)
         i. Name: Name of the coach
         ii. NOC: The country to which he belongs
         iii. Discipline: The sport that the coach is responsible for
         iv. Event: There can exist multiple events for a certain sport. For example, the sport athletics can include Men's 100m, Women's 100m, etc.
      b. The table has 394 instances.

   3. Entries Gender
      a. (Discipline, Female, Male, Total)
         i. Discipline: indicates different sports like athletics, tennis, boxing, basketball, swimming, etc.
         ii. Female: indicates the total number of female participants in that sport.

iii. Male: indicates the total number of male participants in that sport.
iv. Total: is the sum of the female and male column.

b. The table has 46 rows.

4. Medals
   a. (Rank, NOC, Gold, Silver, Bronze, Total, Rank by total)
      i. Rank: is the ranking of the number of gold medals a country won.
      ii. NOC: represents the country.
      iii. Gold: The number of gold medals the country won.
      iv. Silver: The number of silver medals the country won.
      v. Bronze: The number of bronze medals the country won.
      vi. Total: Total number of medals won by the country.
      vii. Rank by total: Rank of the country based on the total number of medals won by the country.
   b. The table has 93 rows.

5. Teams
   a. (Name, Discipline, NOC, Event)
      i. Name: Name of the country.
      ii. Discipline: The sport
      iii. NOC: Duplicate column/ same as Name.
      iv. Event: There can exist multiple events for a certain sport. For example, the sport athletics can include Men's 100m, Women's 100m, etc.
   b. The table has 743 rows.

For the new tables that we aim to create, we plan to retrieve the data by web scraping the olympics website or finding it from another kaggle dataset. The organization and formatting of data into a relational table can be done using a scripting language like Python.

5. Detailed Functionality

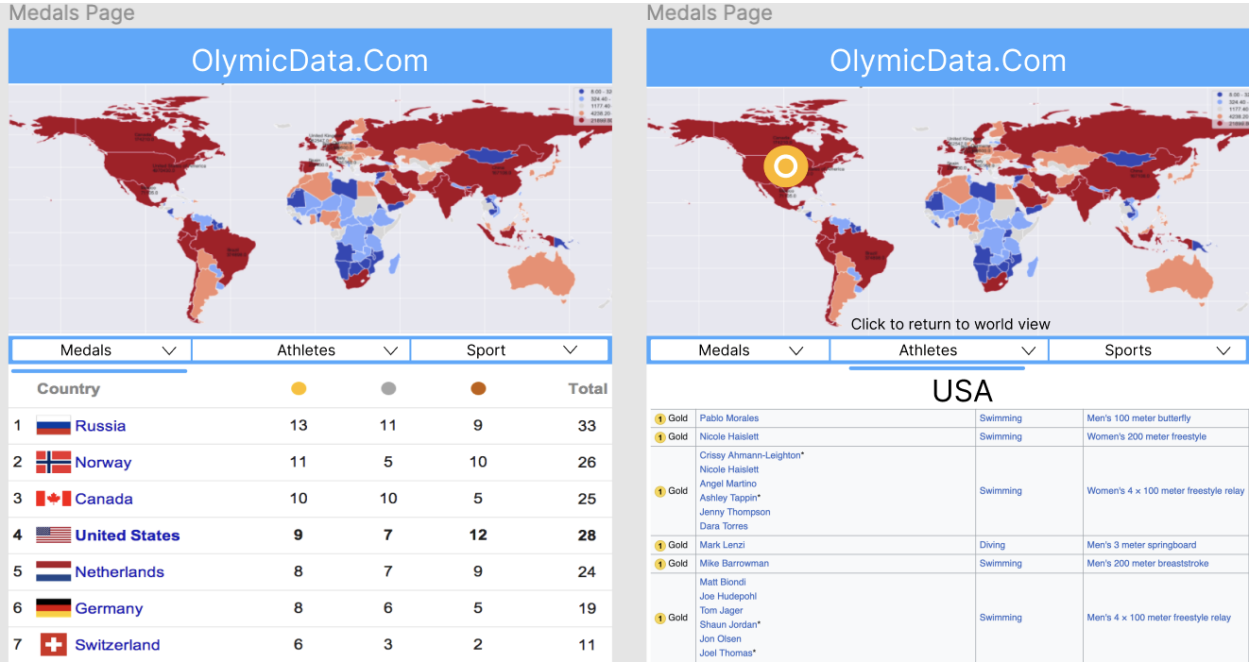
1. The application aims to create tables to include information about
   1. the sport/discipline in which a country secured a medal
   2. the athletes who secured medals for the country.
   3. the events in each sport and the winners of that event.

2. The features of the application also include:
   1. Single/multi attribute filtering by
      a. sport/discipline
      b. event
      c. country

d. athlete
e. coach
f. medals (gold, silver, bronze, total)
g. gender

2. Different button clicks leading to visualizations.
    a. The number of medals secured vs Country.
    b. The number of athletes vs Country.
    c. The number of disciplines each country participated in.
    d. Medals vs Disciplines per country.
    e. Gender ratio in various disciplines.
    f. Coaches per discipline per country.
    g. Number of team sports and individual sports per country.

6. UI Mockup



Filters will come from dropdowns in the selectors directly below the map

7. Work Distribution

1. **Nick**
    a. Query Logic for multi-attribute search. To name a few,
        i. Filter by country and discipline.
        ii. Filter by medals and country.
        iii. Filter by Gender and athlete
    b. Creating overall design and implementation of the front end. Planning on using react for interface

2. **Amrutha**
    a. Web scraping of data for creation of new tables.
        i. Plan to use python's urllib and beautiful soup libraries to achieve the functionality.
    b. Insert/Update and Constraint Logic for creation of new tables.
        i. Country, Sport, Medal
        ii. Country, Athlete, Individual Sport, Medal
        iii. Athlete, Gender
        iv. Event, Gold, Silver, Bronze
    c. Query and visualization Logic for visualization components.
        i. The number of medals secured vs Country - Pie Chart
        ii. The number of athletes vs Country - Bar graph/ Geo chart
        iii. The number of disciplines each country participated in - Bar graph
        iv. Medals vs Disciplines per country - vertical clustered bar graph
        v. Gender ratio in various disciplines - Grouped bar chart
        vi. Coaches per discipline per country - Scatter plot
        vii. Number of team sports and individual sports per country - stacked bar chart
    d. Deployment on GCP.

3. **Lakshay**
    a. Insert/Update and Constraint Logic for creation of new tables.
        i. Country, Sport, Medal
        ii. Country, Athlete, Individual Sport, Medal
        iii. Athlete, Gender
        iv. Event, Gold, Silver, Bronze
    b. Deployment onto GCP
    c. Connecting with front end
    d. Creating query logic for searches for each page