CSC343 Project Phase 1

Steven Yuan, Angelina Zhang
September 30, 2021

Domain: 2016 Olympics in Rio de Janeiro

Datasets:

https://www.kaggle.com/rio2016/olympic-games

This contains 3 datasets, `athletes.csv`, `countries.csv`, and `events.csv` with information about 306 events that took place during the 2016 summer Olympics in Rio as well as 11,517 athletes from 201 countries that participated in them, and the athletes' performance.

The only irrelevant information in these datasets is the *venues* attributes in `events.csv` since we are not interested in the location of the events, we will be more focused on the sport itself.

Investigative Questions:

1. How does a country's living condition indicated by its GDP per capita affect its performance in the Olympics indicated by the number of medals won by its athletes?

2. What does the ideal athlete for each sport look like? That is, does the weight, height, and age of an athlete affect their likelihood to obtain a medal in their sports category? And which category of sport has the strongest such correlation?

3. What is a sporting event (if it exists) that each country is most likely to obtain more medals than other countries in?

Relations:

Athletes(id, name, sex, dob, height, weight)

A tuple in this relation represents an athlete that participated in the 2016 Olympics in Rio de Janeiro. *id* is the unique id of the athlete, *name* is the name of the athlete, *sex* is the sex of the athlete(female, male), *dob* is the date of birth of the athlete(can be considered as age), *height* is the height of the athlete in meters, *weight* is the weight of the athlete in kilograms.

Attribute	Description	Type	Required	Default
id	the identifying number	INT	Yes	

	assigned to each athlete			
name	the first and last name of the athlete	TEXT	Yes	
sex	the sex of the athlete	TEXT	Yes	
dob	the date of birth of the athlete	TIMESTAMP	Yes	
height	the height of the athlete measured in meters	FLOAT	Yes	
weight	the weight of the athlete measured in kilograms	FLOAT	Yes	

Countries(country, <u>code</u>, population, gdp_per_capita)

A tuple in this relation represents a country which includes some information about the economic status of the country when the 2016 Olympics was played. *country* is the name of the country, *code* is the upper-case three-letter country code of the country, *population* is the population of the country, *gdp_per_capita* is the quotient of the country's gdp and its *population*.

Attribute	**Description**	**Type**	**Required**	**Default**
country	the name of the country	TEXT	Yes	
code	the abbreviated code of the country used in the Olympics	TEXT	Yes	
population	the number of people living in the country	INT	Yes	
gdp_per_capita	the gross domestic product per population of the country	FLOAT	Yes	

Performance(<u>id</u>, nationality, sport, gold, silver, bronze)

A tuple in this relation represents the fact that an athlete obtained a certain number of gold, silver, and bronze medals for a particular sport in the 2016 Olympics. *id* is the unique id of the athlete, *sport* is the sport category that the athlete participates in(represented by lowercase strings; its possible values are described in the integrity constraint section below), *nationality is* the nationality of the athlete(represented by an uppercase three-letter country code), *gold* is the

number of gold medals that the athlete obtained for *sport*, *silver* is the number of silver medals that the athlete obtained for *sport*, and *bronze* is the number of bronze medals that the athlete obtained for *sport*.

Attribute	Description	Type	Required	Default
id	the identifying number assigned to each athlete	INT	Yes	
nationality	the country that the athlete is participating for	TEXT	Yes	
sport	the name of the sport that the athlete competed in	TEXT	Yes	
gold	the number of gold medals the athlete obtained	INT	Yes	
silver	the number of silver medals the athlete obtained	INT	Yes	
bronze	the number of bronze medals the athlete obtained	INT	Yes	

Events(<u>id</u>, sport, discipline, name, sex)

A tuple in this relation represents a sport event. *id* is the unique id of the sport event, *sport* is the sport category that the event belongs to, *discipline* is the branch of *sport* that includes multiple events, *name* is the name of the event, and *sex* is the sex of the participants of the event.

Attribute	Description	Type	Required	Default
id	the identifying number of the event	INT	Yes	
sport	the name of the type of sport the event is for (ex. aquatics)	TEXT	Yes	
discipline	the specific discipline of the event (ex. freestyle)	TEXT	Yes	

name	the name of the event	TEXT	Yes	
sex	the sex of all of the event's participants	TEXT	Yes	

Integrity Constraints:

- Athletes[id] ⊆ Performance[id]
- Performance[id] ⊆ Athletes[id]
- Performance[nationality] ⊆ Countries[code]
- Countries[code] ⊆ Performance[nationality]
- Performance[sport] ⊆ Events[sport]
- Events[sport] ⊆Performance[sport]
- Performance[sport] ⊆ {'aquatics', 'archery', 'athletics', 'badminton', 'basketball', 'boxing', 'canoe', 'cycling', 'equestrian', 'fencing', 'football', 'golf', 'gymnastics', 'handball', 'hockey', 'judo', 'modern pentathlon', 'rowing', 'rugby sevens', 'sailing', 'shooting', 'table tennis', 'taekwondo', 'tennis', 'triathlon', 'volleyball', 'weightlifting', 'wrestling'}

Justification of Design:

We are given 3 datasets - **athletes**, **countries**, and **events**, and we kept the structure of **countries** and **events** datasets because the columns of these datasets have one-to-one relationships; in other words, it is not necessary to split them into different relations, and merging them into other relations would cause undesired one-to-many relationships. Also, the **countries** and **events** datasets contain information that is only relevant to the corresponding entities that they represent; for example, it would not make much sense if we were to merge the attributes, *population and gdp_per_capita*, into the **athletes** dataset, as the attributes describe information about a country but not an athlete. For the **athletes** dataset, we divided it into 2 relations - **Athletes** and **Performance**, where **Athletes** contains some biological information about every individual athlete and **Performance** illustrates how an athlete performed for a particular sport. We chose this design because **Athletes** and **Performance** describe different things - we can deduce biological relationships from **Athletes** (such as the relationship between age, height, and weight) and in **Performance**, we can deduce the athletes' performance in a sport category as well as the countries' performance in a sport category. The relations were designed such that there are no redundant attributes or information, as the rows of different relations describe different data and entities, but they are also closely related to one another so that we can extract relationships between different attributes and provide clues to support our investigation.