CSC343 Project Phase 2

Steven Yuan, Angelina Zhang
November 19, 2021

Datasets

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

Design Decisions

Revised question 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? And how does GDP affect the performance?

(We made this change because GDP per capita is GDP divided by population, which is highly influenced by the denominator. But we would also like to see how the overall economic health of a country affects the performance of athletes.)

Revised question 3 (according to TA's suggestion, we break our original question into 2):
In which sport category (if exists), is each country 1) most likely to obtain more medals than other sports and 2) most likely to obtain more medals than other countries?

In phase 1, the TA suggested that we should not only consider relevant attributes, but also the irrelevant attributes. So, we have decided to remove the *sex* attribute in all relations, since none of our research questions require us to extract information based on the *sex* attribute.

Apart from that, in the *Events* relation, we have removed all the attributes except *sport* because in the *Athletes* relation, we only know the medal count for the general *sport* category for each country, not the specific *discipline* or the *name* of the event. Since the *id* is removed, we set *sport* to be the primary key of *Events* so that *Performance(sport)* can have a foreign key constraint that references the *sport* attribute of *Events*.

In addition, we have created a new *Entities* table that stores all the countries/organizations that participated in the 2016 Olympics and added foreign key constraints to *Performance* and *Countries* so that any countries and organizations have to be in the *Entities* table.

Overall, we now have 5 tables, *Athletes, Performance, Countries, Events,* and *Entities*, where *Events* and *Entities* are only used in check constraints to make sure that all the sports are in the *Events* table and all the countries/organizations are in the *Entities* table.

We decided to keep the rest of our schema the same as phase 1 because each table only contains information relevant to the entities that it represents. Each is self contained so that there is no redundant information or attributes, but are closely related enough so that we can extract relationships across tables that aid us in finding answers to our investigative questions.

Cleaning Process

Missing data in *Athletes*: dob: 1, Height: 330 (3%), Weight: 659 (6%)
Missing data in *Countries*: Population: 5 (2%), gdp_per_capita: 25 (12%)

Our first investigative question requires information about a country's GDP and GDP per capita in 2016; since that information is easily accessible and there are only 25 countries with missing data in our dataset(every country with missing population also has missing gdp_per_capita), we decided to fill in the missing data about population and gdp_per_capita by obtaining information from, https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?end=2016&start=2016 and https://data.worldbank.org/indicator/SP.POP.TOTL?end=2016&start=2016.

Even though our second investigative question requires information about individual athletes, unlike information about each country, that information is not accessible, so we decided to remove rows in athletes.csv with missing data. However, this should not have a great impact on our research results, since the percentage of rows with missing data is around 6% (since most rows with missing weight attribute also have missing height attribute).

Note that it is possible for all athletes from a particular country to be missing either weight or height attribute; thus, when cleaning the *Countries* dataset, we removed any country that no longer exists in the athletes table after removing rows with either height or weight attribute. Then, we removed all countries that do not have any athletes in the *Athletes* table since this information is not relevant to our research. A country having no athletes in the *Athletes* table implies that the country did not participate in the Olympics and all of our investigative questions relate to countries and athletes that did participate.

Finally, we converted the *Events* dataset to support "latin-1" encoding so that we were able to successfully import the data into the database without any encoding errors.

When importing data to the database, we first create temporary tables *AthletesTmp* and *EventsTmp* to load the entire dataset and then extract the columns needed(described in the Design Decisions section) from these temporary tables to the actual *Athletes* and *Events* tables, since Postgresql does not allow us to import data to tables with fewer columns than the original dataset.