# Database and Advanced Techniques

Midterms Paper

Student ID: 220516639  
Name: Teo Jia En  
Date of Submission: 5th January 2025

# 1. Critique of Dataset

1.1 Introduction

From record-breaking feats to historic achievements, the Olympics is regarded as the world’s foremost international sports event. [1] With its roots in Ancient Greece and its first modern version in Athens, the Olympics is an age-old event that everyone looks forward to on a biannual basis. This year, Paris Olympics offered unforgettable moments, with Simon Biles winning gold in the women’s vault after her two-year break [2] , to the introduction of breaking, or break-dancing, as a new sport [3]. What makes the Olympics captivating is that athletes from all around the world congregate together to compete on many different sports. [4] This makes cataloguing the event a complex endeavour, as the games and athletes evolve over time. Hence, a profound fascination with the Olympic Games serves as the impetus for this project.

1.2 Dataset

The primary data source for this project is a publicly available historical dataset that documents the past 120 years of Olympic games. It encompasses essential information pertaining to each game, including where it is held, the city it is held at, and various details on the athletes and the medals they won. [5]

The original dataset contains 271,116 entries and 15 columns.

|  |  |  |
| --- | --- | --- |
| 1 | ID | Unique identifier for every athlete |
| 2 | Name | Full name of the athlete |
| 3 | Sex | Denotes whether the athlete is male or female. Since the dataset covers the Games up to 2016 and the first openly non-binary athlete participated in the Tokyo 2020, there are only two possible values for this attribute in this project. [6] |
| 4 | Age | Age of the athlete at that game |
| 5 | Height | Height of the athlete at that game |
| 6 | Weight | Weight of the athlete at that game |
| 7 | Team | Name of the team the athlete represents |
| 8 | NOC | National Olympic Committee 3-letter code |
| 9 | Games | The year and season of the game |
| 10 | Year | The year the game is held in |
| 11 | Season | Winter or Summer |
| 12 | City | The city the game is held in |
| 13 | Sport | Sport the event is in (Eg. Swimming) |
| 14 | Event | Even the athlete is competing in (Eg. Men’s Individual Butterfly) |
| 15 | Medal | Medal the athlete achieved |

Data sources that have been considered, but ultimately eliminated, for this project include:

* Olympics official website: This authoritative website contains a substantial volume of information that are dispersed across numerous web pages. [5] Consolidating this information would require the use of web scrapers. However, the site has no clear policy on the use of automated robots. To prevent any potential legal infractions, this project opted to exclude the use of this website.
* Olympic Data Feed (ODF): This repository contains documentation of past Olympic games and serves as the Documentation Web site of the International Olympic Committee (IOC). [6] However, the information is stored as pdf documents and could not be easily consolidated.

The project focuses on a narrower scope the Summer and Winter Olympics from 2010 to 2016.

The historical data

This dataset is acquired from Kaggle,

where it is uploaded by Randi Griffin

Questions that we would like to ask the dataset

2. Model of data

3. Creation of database

4. Simple web application