# COMS4060A/7056A:

Univeristy of the Witwatersrand 2023

October 5, 2023

Due: October 26, 2023

## 1 Introduction

This assignment is based on content covering geospatial and time series data. You will be required to perform some basic data cleaning and exploration techniques on prescribed datasets. The aim is to explore the dataset and make observations. There is no strict requirement on the format of your submission, but any answers should be reasoned, discussed, and relevant data or results should be provided to substantiate this. You might like to submit either a PDF or a Jupyter notebook, for example. You can use any programming language or tool you would like, however
Submissions should include both text and code. You may use any language that you like.

## 2 The Data Set

The dataset contains a list of all athletes to ever participate in the olympic games.

The athletes.csv file contains a list of athletes who participated at the olympics along with several other variables including. Notably athletes who competed in multiple olympic games or multiple events at the same olympics are presented twice.

- Sex: The athlete's sex. M for Male F for Female.
- Age: The athlete's age at the olympics in years.
- Height: The athlete's height, in cm
- Weight: The athlete's weight, in kg
- Team: The athlete's country
- NOC: The 3-letter abreviation of the athlete's country

- Games: A string combining year and season
- Year: The year in which the olympics took place
- Season: Summer or Winter. Determining if it was a summer or Winter olympics
- City: the host city of the games
- Sport: The sport the athlete competed in
- Event: The particular sub-discipline or category the athlete competed in
- Medal: Gold, Solver. Bronze or NA depending on what medal the athlete won (NA means "no medal").

# 3 Instructions

## 3.1 Removing Outliers and bad data

Find all data-points which seem to be mislabeled. Correct them if possible (google is allowed here). This should include outliers in height or weight who do not seem to be real people, Olympic events that never occured and countries/NOCs that do not exist.

## 3.2 By Country Analysis

Generate a new .csv file of all countries/NOCs called countries.csv. Include columns for:

(a) The number of golds, silvers, bronzes each coutnry has won and the total number of participations (counting athletes who've competed multiple times a corresponding number of times).

(b) Include mean, maximum and minimum weights and heights and fraction of athletes of each gender.

(c) The number of sports each country has medaled and participated in (one variable each).

Use this new csv to:

(a) Use the shape file to visualize how many of each medal type each country has won.

(b) Use the shape file to visualize how many sports each country has won medals in.

(c) Use the shape file to visualize how many sports each country has participated in.

(d) Comment on your results

### 3.3 By Sport Analysis

Generate a new .csv file of all events called events.csv. It should include the following columns (you are welcome to add more).

(a) The number of golds, silvers, bronzes each country has won and the total number of participations (counting athletes who've competed multiple times a corresponding number of times).

(b) Include mean, maximum and minimum weights, heights, and ages.

(c) The number of sports each country has medaled and participated in (one variable each).

Use this new csv to:

(a) Plot the mean weight against the mean height of participants in each sport. Comment on your plot.

(b) Repeat this for weight against age and height against age.

(c) Draw a histogram of how many participations each event has had. Discuss how you choose the number of buckets. Discuss the resulting histogram

### 3.4 Time series Analysis

(a) Create a table called games.csv Each datapoint would by a games. Include the total number of participants per games, the number of unique sports and the number of unique events.

(b) Draw a line graph for each of these variables. Use separate lines for the summer and winter Olympics.

(c) Comment on the results

### 3.5 Other

Do any other data visualisations that you find interesting or menaingful.