**Introduction**

Our last project takes us back to the good old spreadsheet. You will need to extract the data from it before starting your analysis.

We will look at a dataset that contains data about athletes at the Olympic Games.

**The data**

The data is split in four sheets.

* **games**: basic data about the games (season, city...)
* **athletes**: contains basic data about the athletes (name, height, weight, gender...) that are constant for all games. Note that we only have one measure of weight, height and gender so you will have to assume that this did not change even if the athlete participated to the Games more than once
* **country**: contains basic data about the athletes that sometimes changed over time (country, age)
* **medals**: a list of results (type of medals for the winners, NA for the others)

Before you start doing lots of joins, check what happened at the 1956 games, it might be interesting...

**Exercises**

**Part 1**

Have some athletes competed for different countries over time?

**Part 2**

Who are the ten athletes that took part in most games?

**Part 3**

What athlete(s) kept a Gold medal for the longest time?

With keeping a gold medal we mean: competing and winning the gold medal for a *particular event* on the Olympics *several Olympic editions in a row*. E.g. if one wins the gold medal for 100m sprint in 2008 and 2012, you would say they kept this gold medal for two olympic games.

This is a hard question! If you find that you are stuck on this, we would like to point you to a learner webinar that we have given in March 2021 where we explained how you could solve a very similar problem, but on a different dataset. It might give you some useful insights!

The video of this webinar can be found  [EXTS Learning Lab: World's best tennis player (data wrangling) - YouTube](https://www.youtube.com/watch?v=P5cJ8g4fj9A). There is also a github repository with resources attached to this webinar here: <https://github.com/BobbyGlennS/fds-learner-lab>

**Part 4**

What country(ies) kept a Gold medal for the longest time?

**Part 5**

Who are the ten athletes that competed in the most events (some athletes take part in more than one event during games) ?

**Part 6**

Create a new table showing the number of medals per country (rows) and per year (column). Keep only the 15 countries with the most medals overall.

**Part 7**

*Is there a relationship between country and the probability of winning a medal?*

We know that some countries win more medals than others. In this part we want to work out whether some countries win more medals simply because they have more athletes competing for more events, or whether their athletes are in fact performing better in general.

For this part, look only at the top 15 countries that you established in part 6.

* First, create a horizontal barchart listing for each of the fifteen countries the percentage of medals won out of all medals competed for by that country.
* Looking at the chart, what do you think: is there a relationship between country and the probability of winning a medal?
* Run a Chi Square test to test the null-hypothesis: "There is no relationship between country and whether they win a medal or not".
* Include in your report a description of the interpretation of the results. Does it match your expectation based on the bar chart?

**Part 8**

Create a scatterplot showing the average height versus the average weight of competitors per sport (one dot per sport). Add labels with the sport names for:

* the largest average height
* the largest average weight
* the smallest average height
* the smallest average weight
* the largest average BMI
* the smallest average BMI

It might be that the same dot qualifies for multiple labels. The formula to calculate Body Mass Index (BMI) is:

weight in kg / (height in meters)^2

**Part 9**

Create a line plot showing the number of medals given by year (one line for Gold, one line for Silver and one line for Bronze). Does it change over time? Use facets to separate the medals at Summer games and Winter games.