- -> notebooks from this lecture: https://github.com/ine-rmotr-curriculum/data-cleaning-rmotr-freecodecamp
- -> dealing with duplicates
- -> defining what duplicate values are
- -> in this case we have a series which contains ambassadors
- -> we want to invite one ambassador per country
- -> we can see which ones belong to the same country
- -> we need to define the duplicates
- -> political rules for each
- -> duplicated methods returns true
- -> the method is true working top down
- -> for example, if we work our way down in a series and see there are two duplicates
    - -> one country having two ambassadors
    - -> we do this from top down, or bottom up
- -> we can drop duplicates
- -> dropping all the values that check as true
- ***-> for subsets, we have multiple players in the data frame, but they can be in this example present four times***
    - -> so we are understanding the correct subset that we should check
    - -> the season / position that the players played in
    - -> checking for the column name / not checking for it
    - -> which columns we consider as duplicates
    - -> the value and position can be different
- ***-> string handling***
    - -> some types of columns can have specific attributes
    - -> str <- string
    - -> datype columns <- .dt attributes
    - -> .cat attributes
    - -> some have specific methods associated with the domains of the columns
    - -> some of the elements have a lot of methods
    - -> Python can use a split method
    - -> the str attribute in pandas
    - -> splitting the values by an underscore
    - -> then using the expand = True attribute to create a data frame out of this
    - -> contains regular or contains with regular database
    - -> we can fix this with regular expressions
- ***-> question***

The Python method .duplicated() returns a boolean Series for your DataFrame. True is the return value for rows that:

- options
    - contain a duplicate, where the value for the row contains the first occurrence of that value.
    - contain a duplicate, where the value for the row is at least the second occurrence of that value. <- This one
    - contain a duplicate, where the value for the row contains either the first or second occurrence.