In this problem we will try to understand how the gender affects the probability to win a gold medal at the international mathematics olympiad (IMO);

1. Data on participants in international mathematics olympiads (IMO) - folder, IMO_ALL, contains IMO results at the individual level from 1959 to 2018, obtained from https://www.imo-official.org/;

Part 1 - Getting Ready (30 points for each sub-point you get 10 points)

1. Write a **function** which takes as input one of the files in folder IMO_ALL and returns a data.table with the following variables:

(a) Name and the surname of the participant - this variable should be called **name**;

(b) Gender - this variable should be called **gender**;

(c) Country - this variable should be called **country**;

(d) Score of p1 - this variable should be called **p_1**;

(e) Score of p2 - this variable should be called **p_2**;

(f) Score of p3 - this variable should be called **p_3**;

(g) Score of p4 - this variable should be called **p_4**;

(h) Score of p5 - this variable should be called **p_5**;

(i) Score of p6 - this variable should be called **p_6**;

(j) Score of p7 (notice that p7 is present only in some years) - this variable should be called **p_7**;

(k) Total score - this variable should be called **total**;

(l) Rank - this variable should be called **rank**;

(m) Award - this variable should be called **medal**;

(n) Year when the olympiad took place - this variable should be called **year**;

Remove from the data all participants whose name equals * or ?; Also remove all participants who have ? (question mark) in their name (i.e., there is a question mark within a string);

Remark: We will not be checking your code - we will try several random pages and if it works on all these pages you will get a full mark, if not, you will get a grade of 0 for this subpoint.

Hint: It might be helpful to use following functions:

(a) read_html from xml2 package;

(b) Then via xpath find all .//tr nodes;

(c) Then for each .//tr node find all .//td nodes.

2. Write a **function**, which takes as an input, an **absolute path to folder IMO_ALL**, then applies to each .html file in this folder a function from point 1, combines results together rowwise and outputs a .csv file **imo_all.csv**;

   Remark: We will not be checking your code - we will input into your function an absolute path to a folder IMO_ALL and then check the resulting .csv file. If it works you get 5 points, if not 0.

   Christmas gift: imo_all.csv is uploaded on the course platform so you can see whether you get what is necessary;

3. Clean data a little bit:

   You should write a function, function_clean, which does the following:

(a) Takes as an input imo_all.csv (please use fread to read .csv file);

(b) Removes leading "\n" , gender indicator and \n at the end of the name - for example \nBohuslav Diviš ♂\n should now look like, Bohuslav Diviš;

(c) For the variable name, trims white spaces at the beginning and the end of the string;

(d) Creates new variables:

   i. gold_ind =1 if a medal variable contains strings "Gold" or "gold" and 0 otherwise;

   ii. silv_ind=1 if a medal variable contains strings "Silver" or "silver" and 0 otherwise.

   iii. d_female=1 if a participant is a female and 0 otherwise.

   iv. nr_order - the variable showing which time a given participant is participating in an olympiad, participant here is defined by the combination of a country and name;

(e) Replaces empty cells for p1 - p7 with NAs;

(f) Outputs the file "final_data_imo.csv";

Remark: We will not be checking your code - we will input into your function "imo_all.csv" and then check the resulting .csv file. If it works you get 5 points, if not 0.

Christmas gift: final_data_imo.csv is uploaded on the course platform thus you can see whether you get what is necessary;