**Case Study** - Data Analytics Team

Introduction:

Our spreadsheet [here](https://docs.google.com/spreadsheets/d/1Uco1QX6spKv76yAj-e1yCme-kFvRZGG0/edit#gid=312721325) is a collection of data from our events since 2016.

We collect the data during the registration process (either via Zoom or Eventbrite).

The spreadsheet contains two sheets:

1. "events before 052018", unfortunately quality of the data is not really good as in many cases the registrants had the option to fill in what they wanted (instead of pre-filled answers)
2. “events after 052018”, here the quality is way better however not perfect either.

Task:

Your target is to combine these two files and get requested insights from the data.

Task specification:

**Data cleaning** - data in the 2 sheets are a bit mixture of the current questions we are using during the registration and we would need to clean the data to get the same insights for analysis.

Requested data

Age bracket:

15 - 20

21 - 25

26 - 30

31 - 35

36 - 40

41 - 45

46 - 50

51 & above

Education:

Vocational

Bachelor’s

Master’s

Doctorate & higher

Other

Gender

Female

Male

Unspecified

Please write a short description of the process of how you cleaned the data (SQL, Python….)

**Dashboards**

Prepare dashboard based on provided data (choose visualisation tool based on your preference)

We are interested in:

1. Total number of events (completed events)
2. Events per Year
3. Complete number of registrations
4. Unique registrations in total
5. Events per Country (in the first sheet pls work with column - timezone)

* list TOP 5 **countries, add also column with No of registrations**

1. Number of ppl who registered more than once
2. Gender of the registrants
3. Age bracket of the registrants
4. Education
5. You can add any other you consider important