Efficacy analytics audition project – notes

The process of data analysis can be divided into:

1. Uploading data.
2. Checking the quality of data.
3. Analyzing the data using different tools (R, Excel, Tableau 10.1).
4. Making data visualizations.
5. Preparing a report.

My thought process is presented in this document. Visualizations and conclusions are presented in the report file.

1. Uploading data.
   The process of uploading data is not complicated, especially if you have .CSV file. Of course, it differs and depends on the tool you are using.
   In RStudio you type:

```
data1 <- read.csv("C:/Users/Jakub/Desktop/Data.csv", header= TRUE, sep = ",
", dec = ".")
```

   In Excel you open the .CSV file, choose the DATA tab and "Text as columns". Then you need to customize options depending on the data structure.
   In Tableau 10.1 you choose menu Data -> New Data Source -> Text file. Tableau is making connection to the chosen file. You just need to split the text into columns.

2. Checking the quality of data.
   There are 81432 observations of 7 variables (which were mentioned in the instructions).

```
view(data1)
```

   There are no duplicated rows in the data:

```
summary(duplicated(data1))
```

   There are missing values in two columns. Two students do not have **country_id**. Because we have people who are assigned to **country_id**=QU (countries and territories not specified according to ISO classification), I think we can assign people with missing values to this group for our analysis.
   Also two students do not have their **avg_score**. These missing values can be estimated on the basis of the **avg_score** of other learners, comparing people of similar **completion** for this unit. On this basis, I think **NA** values can be replaced by 0.
   Of course, in this case, we had a few missing values and they could be estimated. But it is not always so simple... If there are missing values in the real data, then I think we should also report it to the team which is responsible for the education platform and for collecting the data.
   There are no students assigned to more than one country.

3. Analyzing the data using different tools (and making data visualizations).
   Due to the limited time that can be spend on analyzing, I decided to use mainly Tableau (drag and drop approach; calculated fields). I am still not an advanced user of Tableau, so to do some tasks I had to use other tools.

   First, I checked the number of students in each country. It is worth noting that the number of records is not the same as the number of students!
   Then, I wanted to know:
   - How many students at least started working on particular unit?
   - What is the average of **avg_score**, **completion** and **inv_rate** for each unit?
   - How many students belong to the course taught by the teacher or are studying alone?
   - What are the average of **avg_score**, **completion** and **inv_rate** for these two groups of students?
   - How does this information look like for different countries? Which countries differ significantly from the average?
   - Are there any significant differences in a particular country between average results of students who have and do not have a teacher?

   Also, I need to mention that some of the results are filtered, excluding the countries where the number of students is lower than 20 (the size of one class). Comparing the average for countries with such different number of students can be problematic.

   The tables, charts (and other visualizations) and some findings are available in the report file.