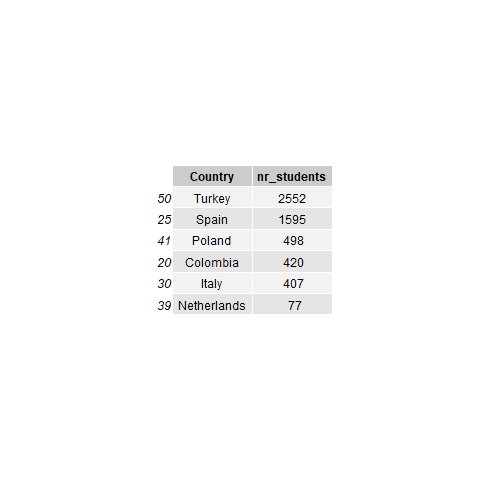
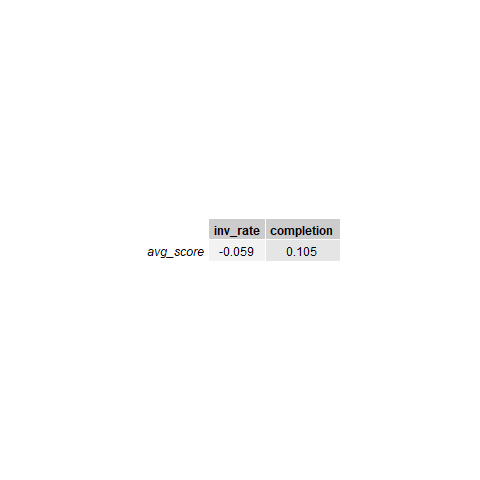
**Data Science Team recruiting task 2018**

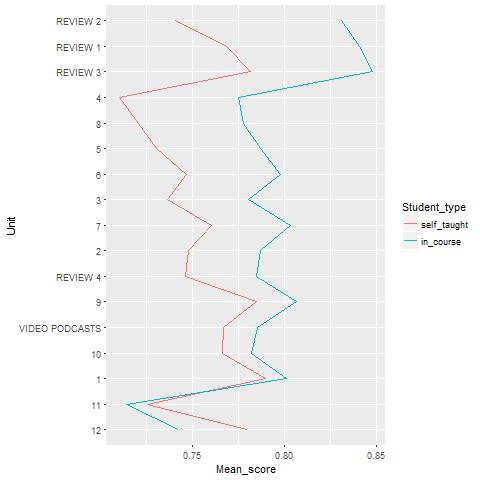
Adrianna Wiśniowska

1. I have decided to clean incoherent data by replacing invalid **avg\_score**, **inv\_rate** and **completion** values by **NA**. I have decided to leave empty values in columns **country** and **unit** in case an empty value bears some meaning.
2. Table below shows 6 countries with the **biggest number of self-taught students**. As those students may need extra assistance in understanding the contents of the book, company should consider creating additional learning materials for students from these countries and/or hiring more support agents speaking national languages of these countries.



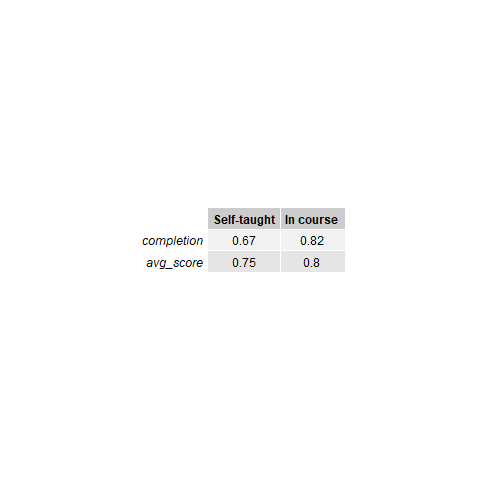
1. Next table shows correlation between **avg\_score** and **inv\_rate** and between **avg\_score** and **completion**. Both correlation values are close to 0, so there is no need to force students to solve exercises in the order proposed by the authors of the book. Low correlation between avg\_score and completion shows that score achieved in a given unit did not affect the moment when student decided to stop solving exercises.



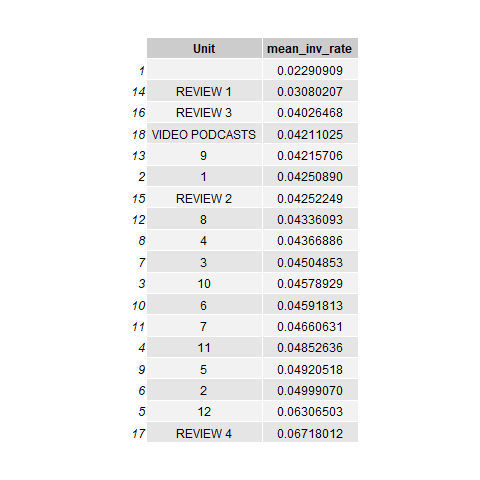
1. Chart below presents an average score of self-taught students and average score of in-course students for every **unit**. Units are ordered by the difference between the average score of self-taught and in-course students.

Self-taught students have had better scores only in **Unit 11** and **Unit 12**. Shapes of both lines are very similar. **Unit 4** is the most difficult for self-taught learners, it is necessary to verify if theoretical part in this unit is clear. Low scores in **Unit 11** may suggest that exercises are too difficult in relation to theoretical part.

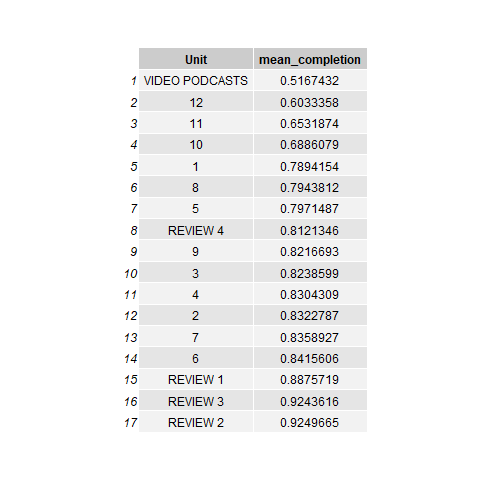
1. The difference in **performance** between self-taught and in-course students is shown below. Their average scores are similar but self-taught students tend to complete less exercises. It can be a sign that some exercises are repetitive or useless from students’ perspective.

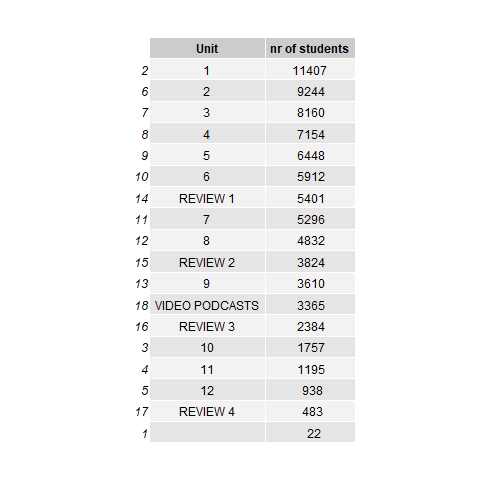


1. All units have small **average inversion rate** (shown below). It is not necessary to reorganize exercises within units, students tend to follow the order imposed by the authors.



1. The majority of units have a high **completion rate**. Video podcasts could be shorter or more diversified in order to attract students.



1. Table below shows chapters and **number of students** who started solving a given chapter. As expected, the further the chapter located in the book, the fewer students start it. If the company wants the students to complete unit **VIDEO PODCASTS** more often, podcasts should either be included in other units or unit should be more prominent in the menu.