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.

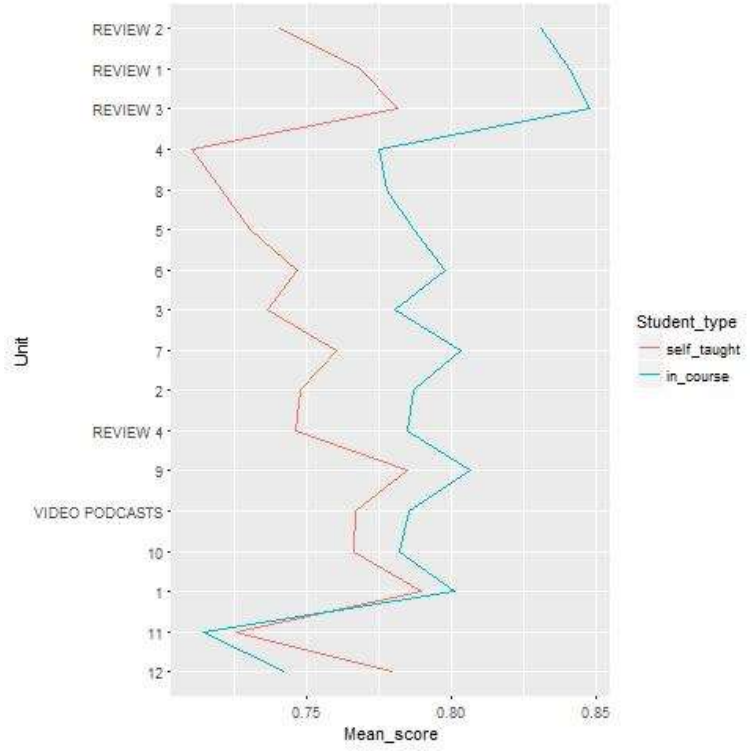
2. Table below shows 6 countries with the **biggest number of self-taught students**. As those students may need extra assistance in understanding the contents of the book, company should consider creating additional learning materials for students from these countries and/or hiring more support agents speaking national languages of these countries.

Country	nr_students
Turkey	2552
Spain	1595
Poland	498
Colombia	420
Italy	407
Netherlands	77

3. 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.

	inv_rate	completion
avg_score	-0.059	0.105

4. 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.

5. 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.

	Self-taught	In course
completion	0.67	0.82
avg_score	0.75	0.8

6. 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.

Unit	mean_inv_rate
	0.02290909
REVIEW 1	0.03080207
REVIEW 3	0.04026468
VIDEO PODCASTS	0.04211025
9	0.04215706
1	0.04250890
REVIEW 2	0.04252249
8	0.04336093
4	0.04366886
3	0.04504853
10	0.04578929
6	0.04591813
7	0.04660631
11	0.04852636
5	0.04920518
2	0.04999070
12	0.06306503
REVIEW 4	0.06718012

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

Unit	mean_completion
VIDEO PODCASTS	0.5167432
12	0.6033358
11	0.6531874
10	0.6886079
1	0.7894154
8	0.7943812
5	0.7971487
REVIEW 4	0.8121346
9	0.8216693
3	0.8238599
4	0.8304309
2	0.8322787
7	0.8358927
6	0.8415606
REVIEW 1	0.8875719
REVIEW 3	0.9243616
REVIEW 2	0.9249665

8. 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.

Unit	nr of students
1	11407
2	9244
3	8160
4	7154
5	6448
6	5912
REVIEW 1	5401
7	5296
8	4832
REVIEW 2	3824
9	3610
VIDEO PODCASTS	3365
REVIEW 3	2384
10	1757
11	1195
12	938
REVIEW 4	483
	22