Data science internship audition project for Pearson IOKI - Report on e-learning platform for English language learners

Jacek Manko

10 02 2018

The following report presents a brief, descriptive analysis of the data set on online language learners. As you can see there were 13157 learners from 87 countries.

```
data_ioki <- read.csv(file = "data2018.csv", sep = ";", na.strings=c("","NA"))

nlevels(as.factor(data_ioki$learner_id))
```

```
## [1] 13157
```

```
nlevels(data_ioki$country)
```

```
## [1] 87
```

A quick look at the frequencies tables revealed that there were striking disproportions in terms of how many units these learners took.

```
sort(decreasing = T, table(data_ioki$country))
```

```
##
##    TR    ES    PL    CO    IT    OM    NL    CH    MX    CZ    AU    RU
## 52710  8409  5948  5537  1779   856   831   554   517   464   432   276
##    TL    CN    RO    BE    HU    UA    QU    BY    SA    GB    NZ    US
##   260   233   218   180   167   159   134   121   111    96    94    94
##    AR    AD    FR    JP    LT    KR    TM    BG    AQ    DE    AZ    EC
##    89    83    78    73    68    58    48    46    42    37    35    29
##    TH    AS    YE    AF    BR    GR    ID    SO    AL    CY    SK    PS
##    29    26    26    24    24    23    23    18    17    17    17    16
##    TN    AX    KW    MD    MK    TC    CR    TW    VN    AG    LV    BL
##    16    14    14    13    13    13    12    12    12    11    11    10
##    HR    MA    VE    VA    XX    CL    CD    GM    ML    AI    DZ    HK
##    10    10    10     9     9     8     7     6     6     5     5     5
##    LY    AN    AO    IQ    PH    CK    IR    SV    SZ    AT    CA    PM
##     5     4     4     4     4     3     3     3     3     2     2     2
##    BD    PT    ZW
##     1     1     1
```

As you can see, students from 9 countries (that took more that 500 units) account for around 90% of all units taken. Therefore, my first concern is to ask if this disproportion was intended or expected and, if not, to undertake some advertising actions to promote this e-learning product in all countries involved in the study. If I narrow the numbers of countries down:

```
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
```

```
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

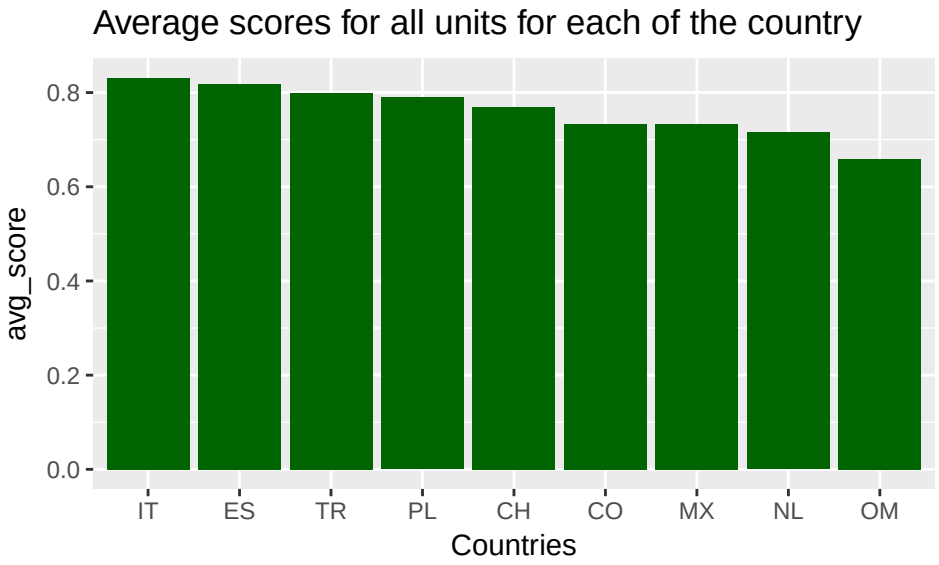
```
data_ioki2 <- data_ioki %>% group_by(country) %>%
  filter(n() >= 500) %>% filter(complete.cases(avg_score, completion))  %>%
  filter(avg_score <= 1)
```

And then calculate the numbers of the learners:

```
data_ioki2 %>% select(learner_id, country) %>%
filter(complete.cases(learner_id, country)) %>%
distinct() %>%
count(country, sort = T)
```

```
## # A tibble: 9 x 2
## # Groups:   country [9]
##   country     n
##    <fctr> <int>
## 1      TR  6661
## 2      ES  1640
## 3      CO  1390
## 4      PL  1262
## 5      OM   325
## 6      NL   273
## 7      IT   252
## 8      CH   208
## 9      MX   106
```

It appears that vast majority of the learners came from Turkey, followed by Spain, Colombia and Poland. Now, let's have a look at differences in how learners from these countries perform:



Interestingly, learners from Mexico, Netherlands and Oman score the worst as compared to other countries. Noteworthy, these are the countries with the smallest numbers of learners. Clearly, there appears to be issue here, with respect to popularity and efficiency of the online workbook. Taking appropriate steps to increase effectiveness of the product in these countries seems to be essential. Italy, however, has on average the best scores, which is even more interesting given small number of learners from that country. It might be important to have a look at who are these learners as well.

Another important issue I'd like to shed some light on is the extent to which all learners make use of the product. Consider the following frequency table:

```
sort(decreasing = T, table(data_ioki$unit))
```

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

As you can clearly see, the number of learners decreases, as the number of unit increases. The last 3 units were taken only by around 11% of the learners. However, since it remains unknown to me, whether all students were given the same time to work with the product, these results are difficult to interpret. It may be that some students didn't simply make it yet to the last chapters, but will do so in the future. Also, some reviews starting from unit 4 are completely missing, even for students who took last chapters, so it might be interesting to have a look why they skip these units. This question becomes more interesting, when you have a look at average scores across all available units:

```
tapply(data_ioki2$avg_score, data_ioki2$unit, mean)
```

```
##             1            10            11            12             2
##     0.7992644     0.7826150     0.7174260     0.7455977     0.7852938
##             3             4             5             6             7
##     0.7793974     0.7731475     0.7851904     0.7965397     0.8021311
##             8             9      REVIEW 1      REVIEW 2      REVIEW 3
##     0.7759762     0.8080241     0.8375481     0.8299643     0.8465100
##      REVIEW 4 VIDEO PODCASTS
##     0.7824045     0.7835875
```

You can see that learners scores on reviews 1,2,3 and 4 slightly better than on the corresponding unit. This is to be expected, since they should be already familiar with the content of the unit at the moment of doing the review. This is why maybe they decide to skip other reviews? Another important point to make, is that average results for chapters 11 and 12 are slightly lower than average score for other units. Given that these chapters are taken least often, it becomes quite interesting to have a look why is that so and what makes those chapters least popular.

At this point, I will address the question what else, except for the country and type of unit, have an impact on average scores? One can assume, it might be the degree of completion, because if one completes more activities within a unit, one should become more proficient at this. However, the data don't substantiate that view clearly.

```
data_ioki2 <- data_ioki %>%
  filter(complete.cases(avg_score, completion))  %>%
  filter(avg_score <= 1)

tapply(data_ioki2$completion, data_ioki2$unit, mean)
```

```
##           1          10          11          12           2
##   0.7894530   0.6889186   0.6531874   0.6033358   0.8322605
##           3           4           5           6           7
##   0.8239502   0.8304072   0.7971239   0.8415606   0.8358927
##           8           9    REVIEW 1    REVIEW 2    REVIEW 3
##   0.7943812   0.8216693   0.8875511   0.9249665   0.9243616
##    REVIEW 4 VIDEO PODCASTS
##   0.8121346   0.5167432
```

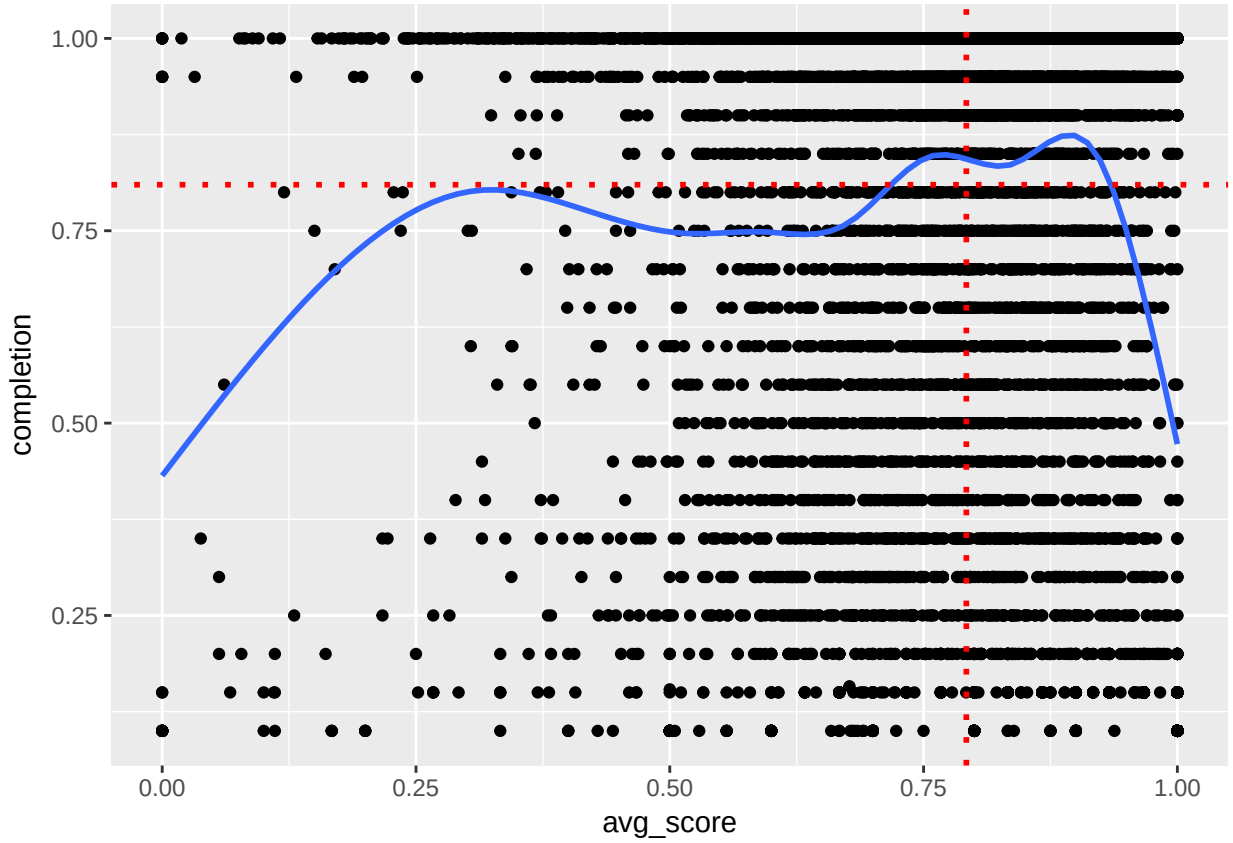
Average completion rates are higher for the reviews. As we already know, average score on reviews were also high. Furthermore, last chapters are characterized by the smallest completion rate. Yet, correlation coeffiecent between average scores and completion rates is only 0.11, which suggests weak positive correlation.

```
cor(x = data_ioki2$avg_score, data_ioki2$completion, method = "pearson")
```

```
## [1] 0.1050119
```

To get a big picture, let's have a look at scatterplot for these variables:

```
## `geom_smooth()` using method = 'gam'
```



For the sake of visibility, this exemplary plot shows only data for the first chapter. Red dotted lines denote means, for both completion and score they are at similar level around 0.78. Blue line denotes relationship between these two variables and you can see this relationship is nonlinear. Although both means are similar, there is no clear relationship between completion rates and average score, because there are some learners with high completion rates, and yet low average score (upper left corner) and many learners with low completion rates, but still, high average scores (lower right corner). Relationship between these two variables would require more detailed look.

Last, but not least, I'd to like to analyze relationship between inversion rates and average scores. One could

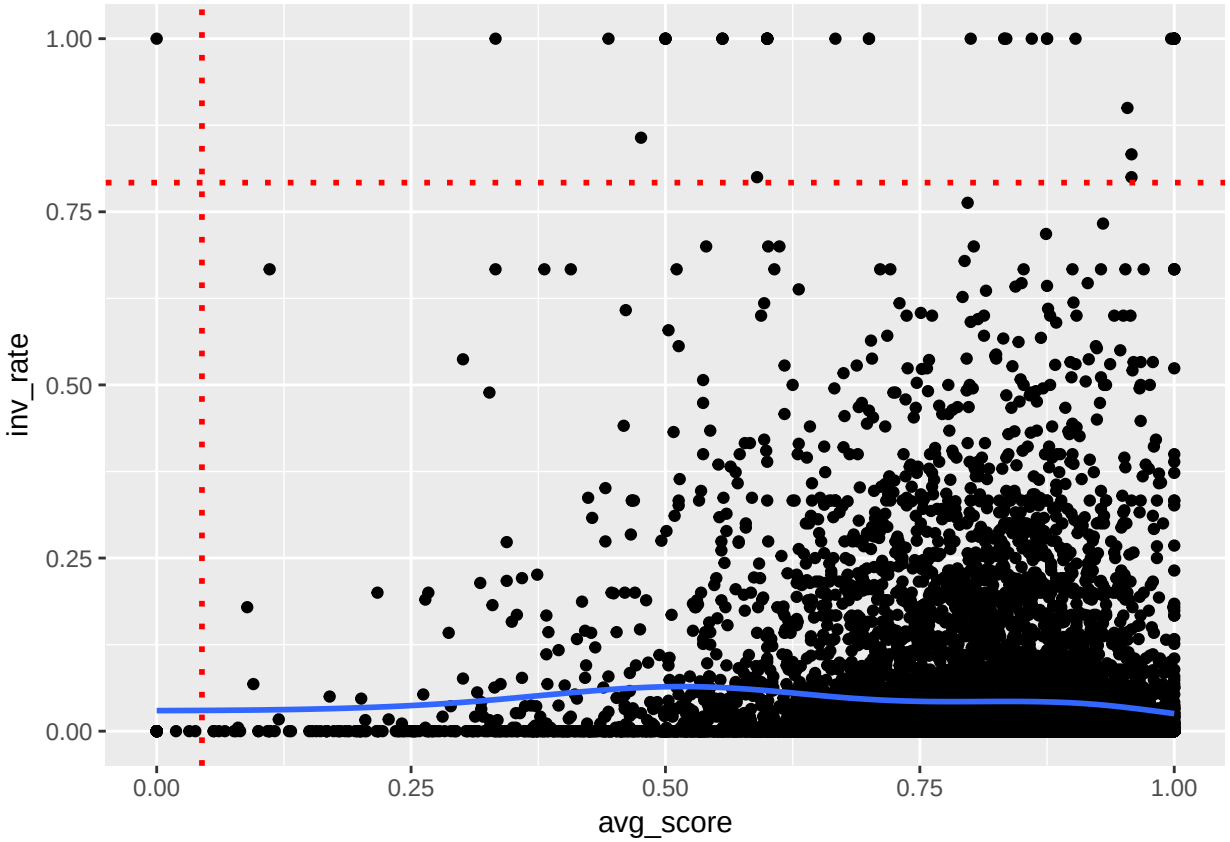
come up with a straightforward hypothesis, that if a learner deviates from the suggested by the experts' order of activities then that learner should receive lower scores. The data suggest no such relationship.

```
data_ioki2 <- data_ioki %>%
  filter(complete.cases(avg_score, inv_rate))  %>%
  filter(avg_score <= 1)
cor(x = data_ioki2$avg_score, data_ioki2$inv_rate, method = "pearson")
```

```
## [1] -0.05927824
```

Although the correlation coefficient is negative, meaning the higher the scores, the lower the inversion rates, the correlation strength is very low, suggesting virtually no relationship between two variables. An analogous plot will make a point more clearly (again, for unit 1 only).

```
## `geom_smooth()` using method = 'gam'
```



Again, although there are many learners in the lower right plot quartile, meaning they had high average scores and low inversion rates, there are other learners with either low inversion rates and low average score or high inversion rates and high average scores. This is quite important finding suggesting that inversion rates have very little, if any, influence on the average scores. Following experts' guidance does not guarantee an optimal outcome on the given unit.

Concluding this short and concise report, this is to say that different students in different countries use this online workbook differently. The data point to some actions that might be taken in order to increase popularity and effectiveness of the product, especially among countries with smallest learner numbers. More detailed analyses, on how exactly learners make use out of the workbook would require more precise data.