Data science internship audition project :: Feb 2017

Task

Show us your data science skills by exploring a dataset from a Pearson e-learning product. Prepare a presentation of your findings and actionable insights that you think would help improve this product.

We're not only interested in the **solution** but also in the **process** you went through to arrive at it. Please send us a Git repository (private or public) that will include both:

- the **presentation** of your findings (in any form you like: slides, markdown, etc);
- all **artifacts** that will show us your **thought process**, e.g. planning notes, code, batch scripts, tests, visualizations, documentation, etc. In other words, we'd like to see anything that can **show us how you approached** this task.

Please submit a link to a public Git repository to your HR contact within 48 hours of receiving these instructions and data.

Note: You can use **any tool** you like to complete this task, e.g. R, Python, Excel, etc. However, since we work mainly in R, you will score bonus points if you show us that you can complete this task using **using R** and such tidyverse packages as tidyr, dplyr, and ggplot2. So even if you don't know R yet, it might be worth spending a weekend to learn the basics using such free sources as this one.

Data dictionary

The dataset comes from a Pearson e-learning platform: an online workbook that is used alongside a paper textbook. Students complete activities online. The activities are either assigned by the teacher as homework or done voluntarily by students as extra practice.

- **student_id**: anonymized student identifier
- **country**: country code of the student
- **in_course**: "t" if the student belongs to course taught by a teacher (as opposed to studying alone)
- **chapter**: number or name of a chapter in the workbook
- **avg_score**: average percentage score on all activities within a given unit
- **completion**: the percentage of activities completed in a given unit, out of all activities available in that unit
- **inv_rate**: This is the extent to which a student deviates from the suggested order of activities by the pedagogy experts within a given unit. A value of zero indicates no departure from the suggested order, a value of one indicates a complete reversal of the order.