Open University Learning Analytics Dataset

Jakub Kuzilek
KMi, The Open University
Milton Keynes, UK
CIIRC, CTU in Prague
Prague, CZ
jakub.kuzilek@open.ac.uk

Martin Hlosta
KMi, The Open University
Milton Keynes, UK
martin.hlosta@open.ac.uk

Zdenek Zdrahal
KMi, The Open University
Milton Keynes, UK
CIIRC, CTU in Prague
Prague, CZ
zdenek.zdrahal@open.ac.uk

Abstract

The purpose of this paper is to present the newly released Open University Learning Analytics Dataset (OULAD). We included the schematic representation of dataset tables, description of data and example manipulation of data using R statistical language. Dataset contains the information about 22 module-presentation, 32593 students, their assessment results and logs of their interactions with the Virtual Learning Environment (VLE) represented by daily summaries of student clicks on different "resource" (10,655,280 entries). Dataset has been anonymized using ARX data anonymization tool.

1 Introduction

In the past decade, the impact of student data analysis has been investigated in over 200 scientific studies [PE14]. They were aiming at the exploration of learning process, especially in e-learning environments. These environments enable universities to collect additional data about student interactions with Virtual Learning Environment, which proved to be a source of useful information, see e.g. [Xen04, AP12, KCdBC15, KHH+15]. It is expected that the universities of this century will be dependent on distant learning students using the VLE. The analysis of student VLE activities will provide the tutors with the information necessary for supporting those who need it [CD15, SAF+14].

In 2012 the Open Academic Analytics Initiative [JML+14] emerged as an initiative aiming at providing Learning Analytics tools to a wider audience and to propose a unified framework for developing such tools.

Finally in 2015 as part of the KDD Conference [KDD] the cup in analysis of student data from XuetangX MOOC courses has been organized. This initiative together with the data released by HarvardX and MITx [HRN+14] in 2014 represents the first effort to propose gold standard data for the Learning Analytics community. This paper describes the dataset provided by the Open University with the aim to

In: A. Wolff, Z. Zdrahal (eds.): Proceedings of the Data literacy for Learning Analytics Workshop, LAK 2016 Conference, Edinburgh, UK, 26-Apr-2016, published at http://ceur-ws.org

contribute to the research in Learning Analytics and to provide another source of data, which can serve the community as the golden standard.

2 Dataset

2.1 Dataset preparation

OULAD is a representative subset of student data collected at the Open University. It consists of student demographics, student performance in course assessments and, last but not least, the log of student behavior in VLE. This data source provides a unique information about student performance and gives the opportunity to create new generations of Learning Management Systems.

As the first step the courses (called modules) with history of at least two consecutive module-presentations were selected. The course represents set of sessions ending with the exam and covering one learning subject. Module-presentation represents one academic year period, in which the module was taught. Data is then transformed and de-identified using the ARX data anonymisation tool [PK15]. After anonymization, the data is checked for errors, certified by the Open Data Institute[1] and released.

2.2 Database scheme

Figure 1 depicts the overall structure of the provided dataset. This dataset is oriented on the students rather then on the course as a whole. Therefore the central table/file is studentInfo, which is connected to courses (one student can have more than one registered course), student_registration, which contains information about dates of registering and deregistering and studentVle, which holds records of student interactions with the VLE system. The structure of the system is captured in table/file vle. Every course contains several assessments (file/table assessment), which are connected with the student using table student_assessment containing record of the student assessments results.

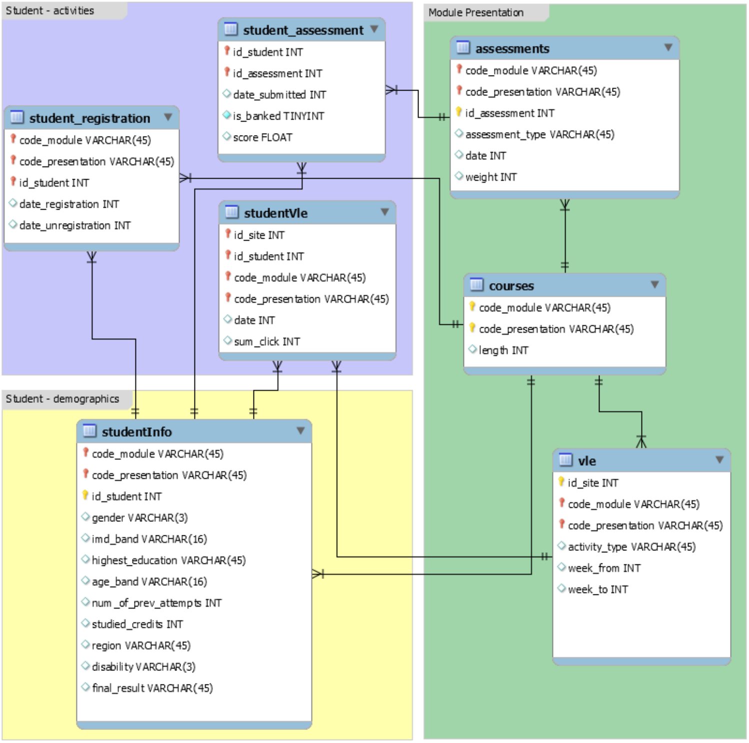


Figure 1: Dataset scheme.

2.3 Data description

2.3.1 courses.csv

File contains the list of all available modules and their presentations. The columns are:

[1] http://theodi.org/

- code_module – code name of the module, which serves as the identifier.
- code_presentation – code name of the presentation. It consists of the year and letter “B” for presentations starting in February and “J” for presentations starting in October.
- length – length of the module-presentation in days.

The structures of B and J presentations may differ and therefore it is recommended to analyse the B and J presentations separately.

2.3.2 assessments.csv

This file contains information about assessments in module-presentations. Usually, every presentation has a number of assessments followed by the final exam. The assessment.csv file contains the following columns:

- code_module – identification code of the module, to which the assessment belongs.
- code_presentation – identification code of the presentation, to which the assessment belongs.
- id_assessment – identification number of the assessment.
- assessment_type – type of assessment. Three types of assessments exist: Tutor Marked Assessment (TMA), Computer Marked Assessment (CMA) and Final Exam (Exam).
- date – information about the final submission date of the assessment calculated as the number of days since the start of the module-presentation. The starting date of the presentation has number 0 (zero).
- weight – weight of the assessment in %. Typically, Exams are treated separately and have the weight 100%; the sum of all other assessments is 100%.

If the information about the final exam date is missing, it is at the end of the last presentation week.

2.3.3 vle.csv

The vle.csv file contains information about the materials available in the VLE. Typically these are html pages, pdf files, etc. Students have access to these materials online and their interactions with the materials are recorded. The vle.csv file contains the following columns:

- id_site – the identification number of the material.
- code_module – the identification code for module.
- code_presentation – the identification code of presentation.
- activity_type – the role associated with the module material. Its context is dependent on module, in which is used. Names of activity types are as much self-explanatory as it can be.
- week_from – the week from which the material is planned to be used.
- week_to – week until which the material is planned to be used.

2.3.4 studentInfo.csv

This file contains demographic information about the students together with their results. Each student might have several rows - each row contains the information about one module student studied. It contains the following columns:

- code_module – the identification code for the module on which the student is registered.
- code_presentation – the identification code of the presentation during which the student is registered on the module.
- id_student – the unique identification number for the student.
- gender – student’s gender.

- region – the geographic region, where the student lived while taking the module-presentation.
- highest_education – the highest student education level on entry to the module presentation.
- imd_band – the Index of Multiple Depravation (UK specific social-economical indicator) band of the place where the student lived during the module-presentation.
- age_band – band of student's age.
- num_of_prev_attempts – the number of how many times the student has attempted this module.
- studied_credits – the total number of credits for the modules the student is currently studying.
- disability – indicates whether the student has declared a disability.
- final_result – student's final result in the module-presentation.

2.3.5 studentRegistration.csv

This file contains information about the time when the student registered for the module presentation. For students who unregistered the date of unregistration is also recorded. File contains five columns:

- code_module – the identification code for the module.
- code_presentation – the identification code of the presentation.
- id_student – the unique identification number for the student.
- date_registration – the date of student's registration for the module presentation. This is the number of days measured to the start of the module-presentation (e.g. the negative value -30 means that the student registered to module presentation 30 days before it started).
- date_unregistration – date of student unregistration from the module presentation, this is the number of days measured relative to the start of the module-presentation. Students, who completed the course have this field empty. Students who unregistered have Withdrawal as the value of the final_result in the studentInfo.csv file.

2.3.6 studentAssessment.csv

This file contains the results of students' assessments. If the student does not submit the assessment, no result is recorded. The final exam submissions is missing, if the result of the assessments is not stored in the system. This file contains the following columns:

- id_assessment – the identification number of the assessment.
- id_student – the unique identification number for the student.
- date_submitted – the date of student submission, measured as the number of days since the start of the module presentation.
- is_banked – the status flag indicating that the assessment result has been transferred from a previous presentation.
- score – the student's score in this assessment. The range is from 0 to 100. The score lower than 40 is interpreted as Fail. The marks are in the range from 0 to 100.

2.3.7 studentVle.csv

The studentVle.csv file contains information about each student's interactions with the materials in the VLE. This file contains the following columns:

- code_module – the identification code for a module.
- code_presentation – the identification code of the module presentation.
- id_student – the unique identification number for the student.
- id_site – the identification number for the VLE material.

- date – the date of student’s interaction with the material measured as the number of days since the start of the module-presentation.

- sum_click – the number of times the student interacts with the material in that day.

3 Example usage

This section provides guide how to use the OULAD data. The following parts will take you through the environment setup, data preparation and data manipulation. Analysis is performed on the selected subset of data. This example was designed for readers to get the "feeling" of data and how they are organised.

3.1 Environment setup

First of all you need to download and install **R version 3.2.2** and **RStudio**. After installing required the software, we need to install package `data.table`, which provides enhanced functionality for the `data.frame` data type in R, by executing this command:

```
install.packages("data.table")
```

After installing, we need to load library `data.table` into the environment. This can be done by executing this command:

```
library(data.table)
```

3.2 Example task

Compare average result of registered female students, who engaged with Virtual Learning Environment in 28th day of presentation, from two different presentations in 1st and 2nd Tutor Marked Assigment (TMA) using combined weighted score.

3.3 Data preparation

First of all we need to download the data by executing this command:

```
download.file("http://kmi-web29.open.ac.uk:8080/resources/documents/mashupData.RData",
              destfile = "./mashupData.RData",
              mode = "wb",quiet = TRUE)
```

In the next step we will load data into the R environment using:

```
load("mashupData.RData")
```

You can observe loaded data in top right corner of RStudio (see Figure 2).

3.4 Solution

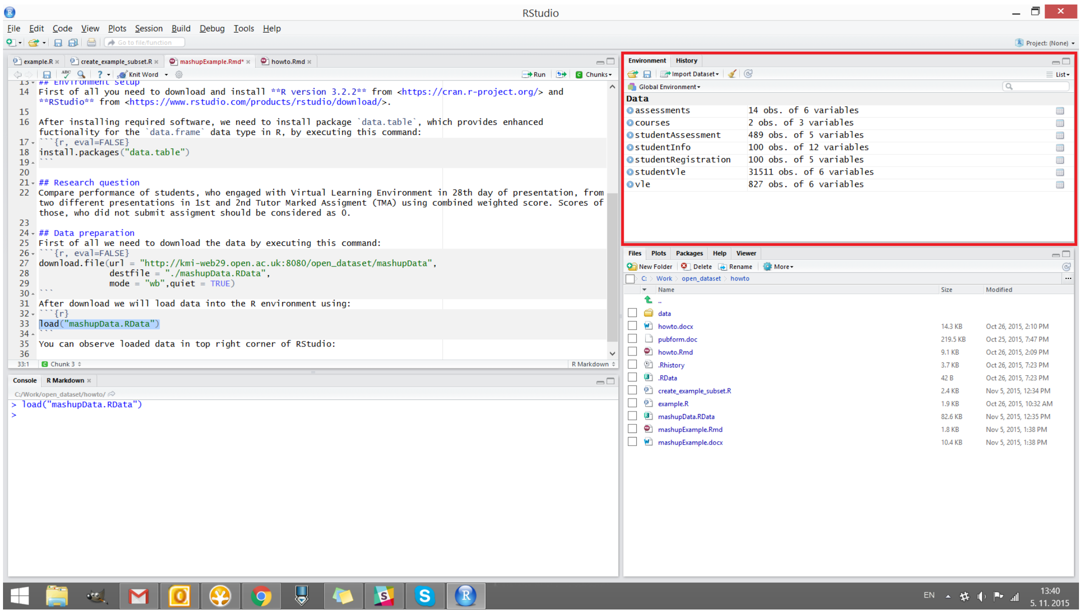
Now, the environment and data are ready and we can start solving the *Example Task*.

3.4.1 Selecting students

First we need to select female students only. The demographic information about students is presented in table `studentInfo`, thus we will apply the following command to select female students:

```
femaleStudents <- studentInfo[gender == "F"]
```

The result of selection will be stored in new variable `femaleStudents` containing demographics of all female students in both presentations.



~~~
## Environment setup
First of all you need to download and install **R version 3.2.2** from <https://cran.r-project.org/> and
**RStudio** from <https://www.rstudio.com/products/rstudio/download/>.

After installing required software, we need to install package `data.table`, which provides enhanced
fuctionality for the `data.frame` data type in R, by executing this command:
```{r, eval=FALSE}
install.packages("data.table")
~~~

Research question
Compare performance of students, who engaged with Virtual Learning Environment in 28th day of presentation, from
two different presentations in 1st and 2nd Tutor Marked Assigment (TMA) using combined weighted score. Scores of
those, who did not submit assigment should be considered as 0.

Data preparation
First of all we need to download the data by executing this command:
```{r, eval=FALSE}
download.file(url = "http://kmi-web29.open.ac.uk:8080/open_dataset/mashupData",
          destfile = "./mashupData.RData",
          mode = "wb",quiet = TRUE)
```
After download we will load data into the R environment using:
```{r}
load("mashupData.RData")
```
You can observe loaded data in top right corner of RStudio:
```

```
C:/Work/open_dataset/howto/
> load("mashupData.RData")
>
```

Data

assessments	14 obs. of 6 variables
courses	2 obs. of 3 variables
studentAssessment	489 obs. of 5 variables
studentInfo	100 obs. of 12 variables
studentRegistration	100 obs. of 5 variables
studentVle	31511 obs. of 6 variables
vle	827 obs. of 6 variables

C: > Work > open_dataset > howto

Name	Size	Modified
..		
data		
howto.docx	14.3 KB	Oct 26, 2015, 2:10 PM
pubform.doc	219.5 KB	Oct 25, 2015, 7:47 PM
howto.Rmd	9.1 KB	Oct 26, 2015, 2:09 PM
.Rhistory	3.7 KB	Oct 26, 2015, 7:23 PM
.RData	42 B	Oct 26, 2015, 7:23 PM
create_example_subset.R	2.4 KB	Nov 5, 2015, 12:34 PM
example.R	1.9 KB	Oct 26, 2015, 10:32 AM
mashupData.RData	82.6 KB	Nov 5, 2015, 12:35 PM
mashupExample.Rmd	1.8 KB	Nov 5, 2015, 1:38 PM
mashupExample.docx	10.4 KB	Nov 5, 2015, 1:38 PM

Figure 2: Dataset scheme.

3.4.2 Selecting registered students

We are interested only in students, who studied at 28th day of presentation, thus we need to filter out students who unregistered from the module before and during this day. Registration data are stored in table `studentRegistration` and we will select only registered students executing this command:

```
registeredStudents <- studentRegistration[
                            is.na(date_unregistration) | date_unregistration > 28
                            ]
```

Note that students, who finished the course has value `NA` in the `date_unregistration` column, thus we need to select students with `NA` (finished the module) or unregistered later than 28th day of presentation.

3.4.3 Selecting VLE active students

We are also interested in students, who were active in VLE at 28th day of presentation. Thus we need the information contained in table `studentVle`. We need to select log entries with `date` column equal exactly to 28. Because we do not need more information than student identification – `id_student`, and which presentation he/she studied we will extract only these columns and unique rows by executing following commands:

```
activeStudents <- studentVle[date == 28][,
                                          .(id_student,
                                            code_module,
                                            code_presentation)
                                          ]

setkey(activeStudents, id_student, code_module, code_presentation)

activeStudents <- unique(activeStudents)
```

3.4.4 Combining student scores from first and second TMA

The required data subsets are ready for solving the task. The next step is the selection of the first and second TMA from each presentation, retrieve their weights and calculate the combined score for each student. Assessment weight is stored in table `assessments` and we need to extract it from this table. We will first extract the `id_assessment` values of all four assessments required to solve our task. This can be done by executing these commands:
```

```
assessments_codes <- assessments[assessment_type == "TMA"]
assessments_codes <- assessments_codes[order(date)][1:4, id_assessment]
```

We first select only TMA type of assessments, then table is reordered according to cutoff date of assessments (noting that first and second assessments has cutoff dates lower than third in any presentation) and then we extract first four ids from the reordered table, which represents the ids of first and second TMA in both presentations. Next we will use `assessments_codes` in extracting all necessary information from `assessments` table executing this command:

```
selectedAssessments <- assessments[id_assessment %in% assessments_codes]
```

Now the required information about assessments is ready and we have to combine it with the student results in `studentAssessment` table. This can be done by setting primary keys for `selectedAssessments` and `studentAssessments` and then joining them together leaving only first and second TMA results, executing this command:

```
setkey(selectedAssessments, id_assessment)
setkey(studentAssessment, id_assessment)

studentAssessmentWithWeights <- studentAssessment[selectedAssessments]
```

Finally, we need to combine the TMA results with the corresponding weights by executing this command:

```
studentAssessmentWithWeights[, weightedScore := score*weight/100]
```

And then for each student sum weighted score together by executing:

```
studentResults <- studentAssessmentWithWeights[,
                                               .(score = sum(weightedScore)),
                                               by=.(id_student,
                                                    code_module,
                                                    code_presentation)
                                               ]
```

3.4.5 Putting everything together

Finally, everything we need is ready and we can process the partial answers. First we will combine `femaleStudents` with `registeredStudents` to get those female students, who is still progressing in module at 28th day of presentation, by executing this command:

```
setkey(femaleStudents, id_student, code_module, code_presentation)
setkey(registeredStudents, id_student, code_module, code_presentation)

registeredFemaleStudents <- femaleStudents[registeredStudents, nomatch=0]
```

Next we will combine the newly created table with table `activeStudents` to select those, who were active in 28th day of presentation, executing this command:

```
activeRegisteredFemaleStudents <-
                registeredFemaleStudents[activeStudents, nomatch=0]
```

And the last step is adding the score for each selected student by executing this command:

```
activeRegisteredFemaleStudentsWithScore <-
                activeRegisteredFemaleStudents[studentResults, nomatch = 0]
```

3.4.6 Comparison of student results from different presentations

Finally we can answer the original question, by executing the following command:

```
print(activeRegisteredFemaleStudentsWithScore[,
                                              .(mean.score = mean(score)),
                                              by=code_presentation])
```

And the result is:

```
##    code_presentation mean.score
## 1:             2013J     16.155
## 2:             2014J      7.500
```

4 Conclusion

Dataset described in this paper is released with the aim to strengthen the Learning Analytics research across the sector of Higher Education and to provide researchers with a useful golden standard data for their experiments. The dataset contains 22 modules with over 30,000 students accompanied by the log of their VLE activities. The dataset has been certified with the Open Data Institute pilot certificate.

4.0.1 Acknowledgements

Our thanks to the Open University for giving us the opportunity of releasing this dataset. Especially to Professor John Domingue, Director of KMi and Professor Belinda Tynan, Pro-vice Chancellor of the Open University - without their support this dataset would never existed.

References

[AP12] Kimberly E Arnold and Matthew D Pistilli. Course Signals at Purdue: Using Learning Analytics to Increase Student Success. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, pages 267–270, 2012.

[CD15] Michael M. Crow and William B. Dabars. *Designing the New American University*. Johns Hopkins University Press, 2015.

[HRN+14] Andrew Dean Ho, Justin Reich, Sergiy O Nesterko, Daniel Thomas Seaton, Tommy Mullaney, Jim Waldo, and Isaac Chuang. HarvardX and MITx: The First Year of Open Online Courses, Fall 2012-Summer 2013. *SSRN Electronic Journal*, (1):1–33, 2014.

[JML+14] Sandeep M. Jayaprakash, Erik W. Moody, Eitel J. M. Lauria, James R. Regan, and Joshua D. Baron. Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. *Journal of Learning Analytics*, 1(1):6–47, 2014.

[KCdBC15] Gregor Kennedy, Carleton Coffrin, Paula de Barba, and Linda Corrin. Predicting success: how learners' prior knowledge, skills and activities predict MOOC performance. *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge - LAK '15*, pages 136–140, 2015.

[KDD] Kdd cup 2015 - predicting dropouts in mooc. http://kddcup2015.com/information.html. Accessed: 2016-01-13.

[KHH+15] Jakub Kuzilek, Martin Hlosta, Drahomira Herrmannova, Zdenek Zdrahal, Jonas Vaclavek, and Annika Wolff. LAK15 Case Study 1: OU Analyse: Analysing At-Risk Students at The Open University - Learning Analytics Review. *Learning Analytics Review*, 2015.

[PE14] Zacharoula Papamitsiou and Anastasios A Economides. Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence. *Educational Technology & Society*, 17(4):49–64, 2014.

[PK15] Fabian Prasser and Florian Kohlmayer. *Medical Data Privacy Handbook*, chapter Putting Statistical Disclosure Control into Practice: The ARX Data Anonymization Tool, pages 111–148. Springer International Publishing, Cham, 2015.

[SAF$^+$14] Mike Sharples, Anne Adams, Rebecca Ferguson, Mark Gaved, Patrick McAndrew, Bart Rienties, Martin Weller, and Denise Whitelock. Innovating pedagogy 2014. *The Open University*, 2014.

[Xen04] Michalis Xenos. Prediction and assessment of student behaviour in open and distance education in computers using Bayesian networks. *Computers and Education*, 43(4):345–359, 2004.