# CSC8631 Coursework Assignment

**Mariela Ayu Prasetyo (210407835)** *Newcastle University*

## Introduction

This report aims to discuss and report any findings to the CSC8631 coursework assignment regarding exploratory data analysis in learning analytics. Regarding the data set, the students were provided with one from FutureLearn MOOC. We are then asked to provide any valuable or non-valuable insights while following the best-practice development explained thoroughly via the teaching resources available. To ensure we are following the data-driven process throughout the project, we will also be adhering to the CRISP-DM methodology. Lastly, we will wrap up with a conclusion regarding the overall findings.

## Business Understanding

The first step of CRISP-DM is the *Business Understanding* step, where we try to understand what the business wants to solve. In this particular project, we are given the data set regarding a course in the FutureLearn MOOC platform. Just like a real-life face-to-face class, we want to find out whether students are involved in the class or not and how many students continue to participate until the end of the semester. We are also interested in finding out how many students leave the class and, naturally, their reason for doing so. Lastly, we want our course to improve constantly, so we will consider any feedback from the data (CHECK). Formulating all of the previous problems into a sentence, we come up with the following precise questions :

- What is the participation rate of the students? (CHECK): and Are there any variables closely related to this?
- How many students and what causes students to leave the course?
- Which part of the learning course can we further improve?

## Data Understanding and Data Preparation

After formulating the questions, we move on to the *Data Understanding* and *Data Preparation* part, where we focus on understanding and formatting the data that assist the business tasks defined in *Business Understanding*. This phase will consist of:

- Describe data: In this part, we are trying to understand and describe the data in a short description. We can do this by examining the data format, the number of rows and columns, and the features that are accessible.
- Exploring the data: In this section, we are trying to analyse the relationship between data and visualise the data. The conclusion and visualisation of the data exploration should support and verify the business question defined previously. We will also select, clean, integrate and formatting the data for easier visualisation and modeling (CHECK)

*Describe Data*

Firstly, we are given the data set regarding an online course from the FutureLearn MOOC platform. The data set consists of several files for each run from run 1 to run 7. Each run may consist of the following data in the .csv form (the number inside the bracket denotes how many variables are there in the file):

- archetype survey response (4)
- enrolments (13)
- leaving survey response (8)
- question response (10)
- step activity (6)
- team members (5)
- video stats (28)
- weekly sentiment survey response (4)

Each of the files consists of different rows (entries), and all the columns (variables) are stored in a chr format.

*Exploring the Data*

In this section, we will begin to explore the given data set. In particular, we want to explore the area where the solution will support the problem defined in the *Business Understanding* part. There are three questions and we will explore them one by one.

*Question 1*

**What is the participation rate of the students? (CHECK): and Are there any variables closely related to this?**

For the first question, we want to analyse whether students are highly engaged in the course. There are many ways to check this, but we will focus on how many percentages of students fully participated and finished the material of the classes. Additionally, we will also check the duration of the completion for each student.
We can do this by checking the enrolments.csv provided in the data set. First, we will do some data pre-processing.

```
#read file enrolments run 1 to 7
files = list.files(path = "data/",
                   pattern="*enrolments.csv", full.names = T)

#store each run in a single variable
for (i in 1:length(files)) {
  temp <- paste("enrolments", i, sep = "")
  assign(temp, read.csv(files[i]))
}
```

We read and store each run of enrolments in a single variable for a later use. After that, we begin to do some processing in our data. We begin by subsetting the students who fully participated in the course (there are entries for the fully_participated columns in the csv data).

```
enrolments <-
  enrolments[!enrolments$fully_participated_at == "" , ]
```

Then, we convert the enrolled_at and fully_participated variables of each run from string format to date format to calculate the duration later on. We convert it by using the as.Date function.

```
as.Date(as.character(enrolments_na$enrolled_at),
        format = "%Y-%m-%d")
as.Date(as.character(enrolments_na$fully_participated_at),
        format = "%Y-%m-%d")
```

After converting it, we calculate the duration between the fully_participated_at and enrolled_at and store it in a new variable. We also remove entries or outliers for each run where the period exceeds 365 days or one year.

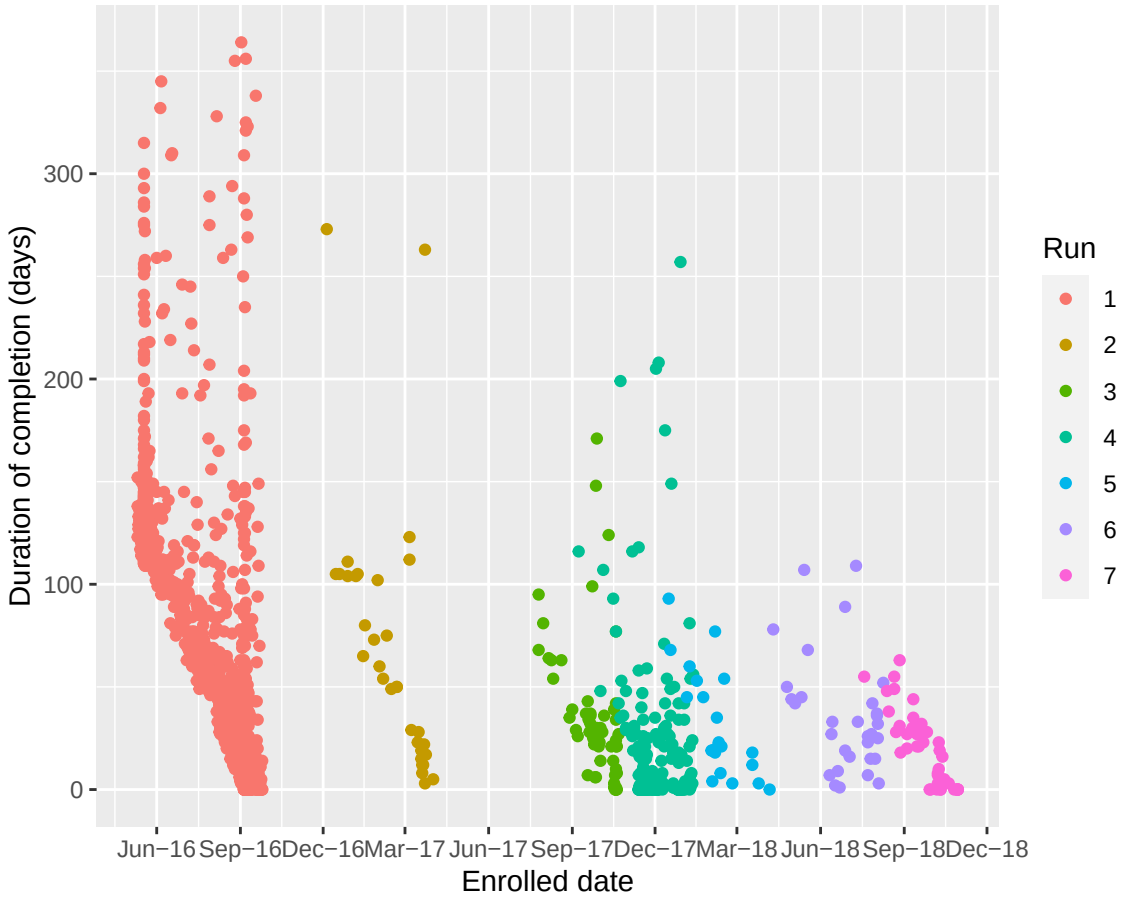
```
enrolments1_na <- enrolments1_na[!(enrolments1_na$duration
                                   > 365) , ]
```

Lastly, we calculate the percentage of students who fully participated in the course proportion to the overall enrollment for each run.

```
enrolments_completion_rate[1] =
  dim(enrolments1_na)[1]/dim(enrolments1)[1] * 100
```

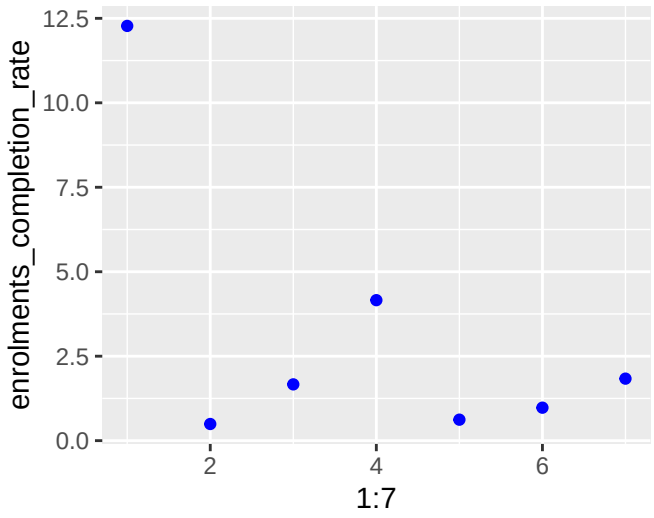
After preparing all of the data, we can now do the graphing. For the first graph, we will graph using ggplot2 the duration of completion against the starting date for each student in each run.

```
# Insert plot
graph1
```




For the second graph, we will include the percentage of students who fully participated in the course proportion to the overall enrollment for each run.

```
# Insert plot
graph2
```



*Discussion*

In graph 1, we can see that duration of students completing the courses vary, but mostly concentrated in 0 to 150 days

*Question 2*

**How many students and what causes students to leave the course?**

*Question 3*

**Which part of the learning course can we further improve?**

*Additional notes*

Configuration setting, when