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 course and, naturally, their reason for doing so. Next, we want to find out which topics students are most interested to learn about. Lastly, we want to know the students' opinion of the course and what to improve in the future. After formulating the previous problems into a sentence, we come up with the following precise questions :

- Is the students in the course highly engaged? (CHECK): and Are there any variables closely related to the participation or engagement rate?
- How many students and what causes students to leave the course?
- What topic interests students the most?
- How are the students who have taken the course feedback?

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 tailor the data by selecting, cleaning, integrating, and formatting it.

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 check the full participation rate. We will focus on how many percentages of students fully participated and finished the material of the classes. We will also check the duration of the completion for each student and we can do this by checking the enrolments.csv provided in the data set.

Firstly, we will do some data pre-processing for part 1

```
#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 using difftime function. We convert it by using the as.Date function. It is also worth mentioning that we store the code inside a function for a more effective writing process for the converting and calculating duration part.

```
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")
difftime(enrolments_na$fully_participated_at,
         enrolments_na$enrolled_at,
         units = c("days"))
```

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.

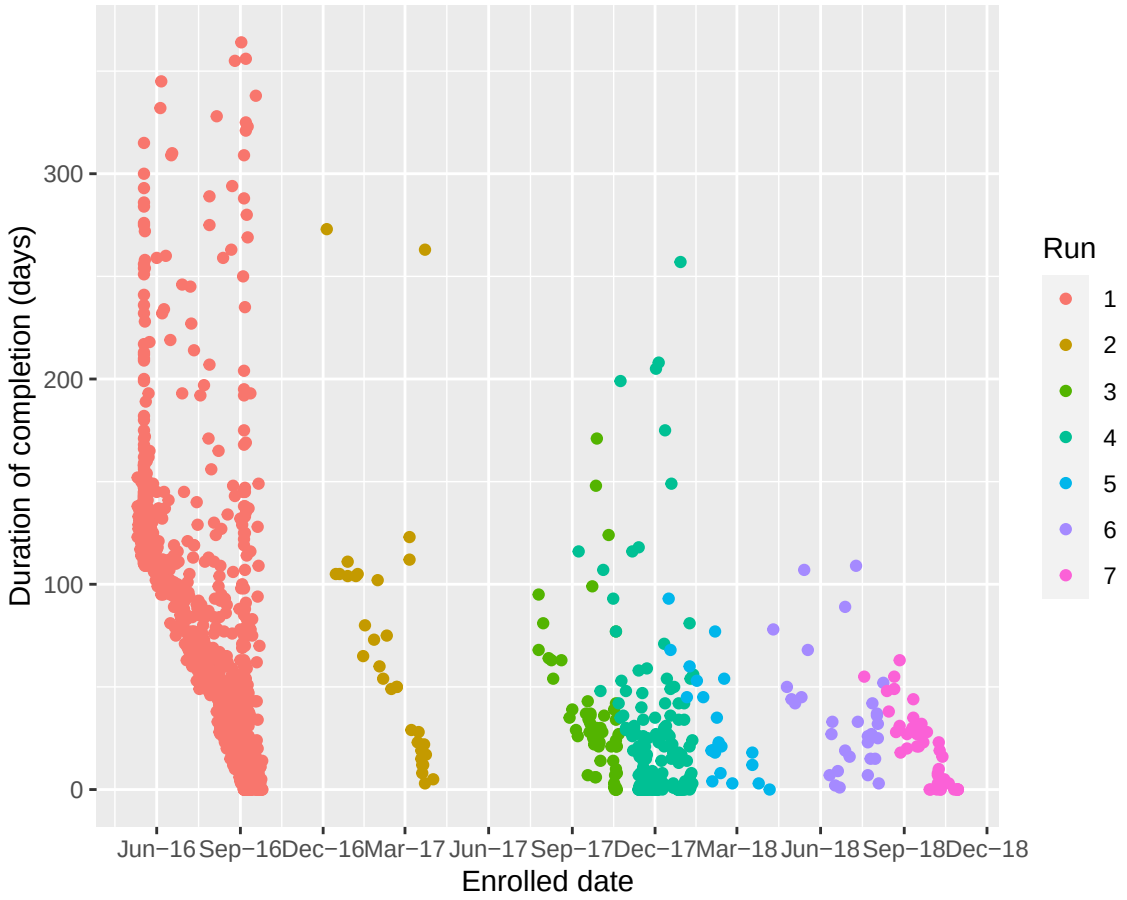
```
enrolments1_na <- enrolments1_na[!(enrolments1_na$duration
                                   > 365) , ]
```

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
```

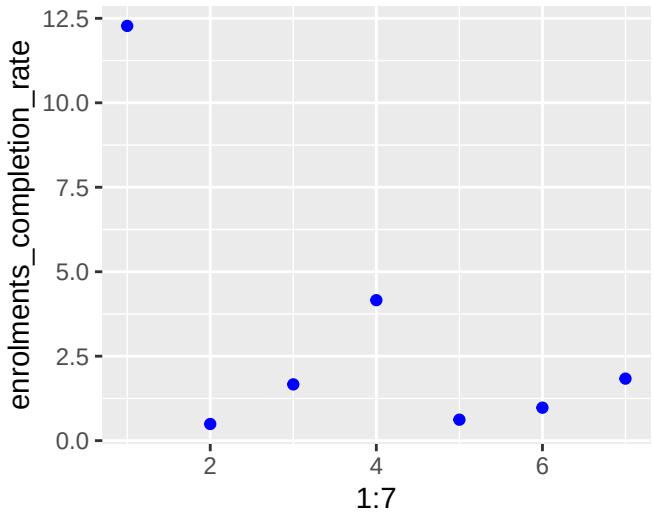
After preparing all of the data, we can now do the graphing. For the first graph, we will graph using ggplot2 the duration of completion against the starting date for each student in each run.

```
# Insert plot 1
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 2
graph2
```



Discussion

In graph 1, we can see that duration of students completing the courses varies but is primarily concentrated in 0 to 150 days. We can also see that the number of people who fully participated in the course decreases for each run. This means that the class gets less engaging as time goes by, and different approaches are needed to boost the engagement rate of the students. Additionally, the graph shows runs that open in the latter half of the year gain more enrollment than those that begin in the first half of the year. This pattern could be taken into consideration when opening a new run in the future.

Graph 2 shows the percentage of students who fully participated in proportion to the overall enrolments against each run. We can see that it peaked at 12.5% for run one and stays under 5% after that. It confirms the statement before that the students who fully participated decreased for each run.

Question 2

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

For the second question, we want to analyse how many students leave the course, what time they quit, and why. Same as before, we will begin by pre-processing the data from leaving survey responses. It is worth mentioning that we assume people who formally quit the course are all required to fill the survey, so the number of students who formally leave the class is equal to the number of the survey entries.

```
#read file archetype survey responses run 1 to 7
#and store it in a variable
files = list.files(path = "data/",
                   pattern="*leaving-survey-responses.csv"
                   , full.names = T)

for (i in 1:length(files)) {
  temp <- paste("leaving", i, sep = "")
  assign(temp, read.csv(files[i]))
}
```

We read the file and store each run data in a variable similar to before. We can then move on to the processing part. We begin by merging the enrolment and leaving data to extract the enrolment_at column. The column later will be used to calculate the duration between the starting and leaving date.

```
leaving4 <- merge(leaving4, enrolments4, by = "learner_id")
```

Then, we drop the columns unrelated to our observations, such as id, last_completed_step_at, etc.

```
leaving4 <- leaving4[-c(5:8, 10:20)]
```

Like before, we convert the date from string format to date format and calculate the duration between the starting and leaving dates using difftime function.

```
as.Date(as.character(leaving$enrolled_at),
        format = "%Y-%m-%d")
as.Date(as.character(leaving$left_at),
        format = "%Y-%m-%d")
difftime(leaving$left_at, leaving$enrolled_at,
         units = c("days"))
```

We also compute how many percentages of students formally quit the course.

```
leaving_num[1] = dim(leaving4)[1]/dim(enrolments4)[1] * 100
```

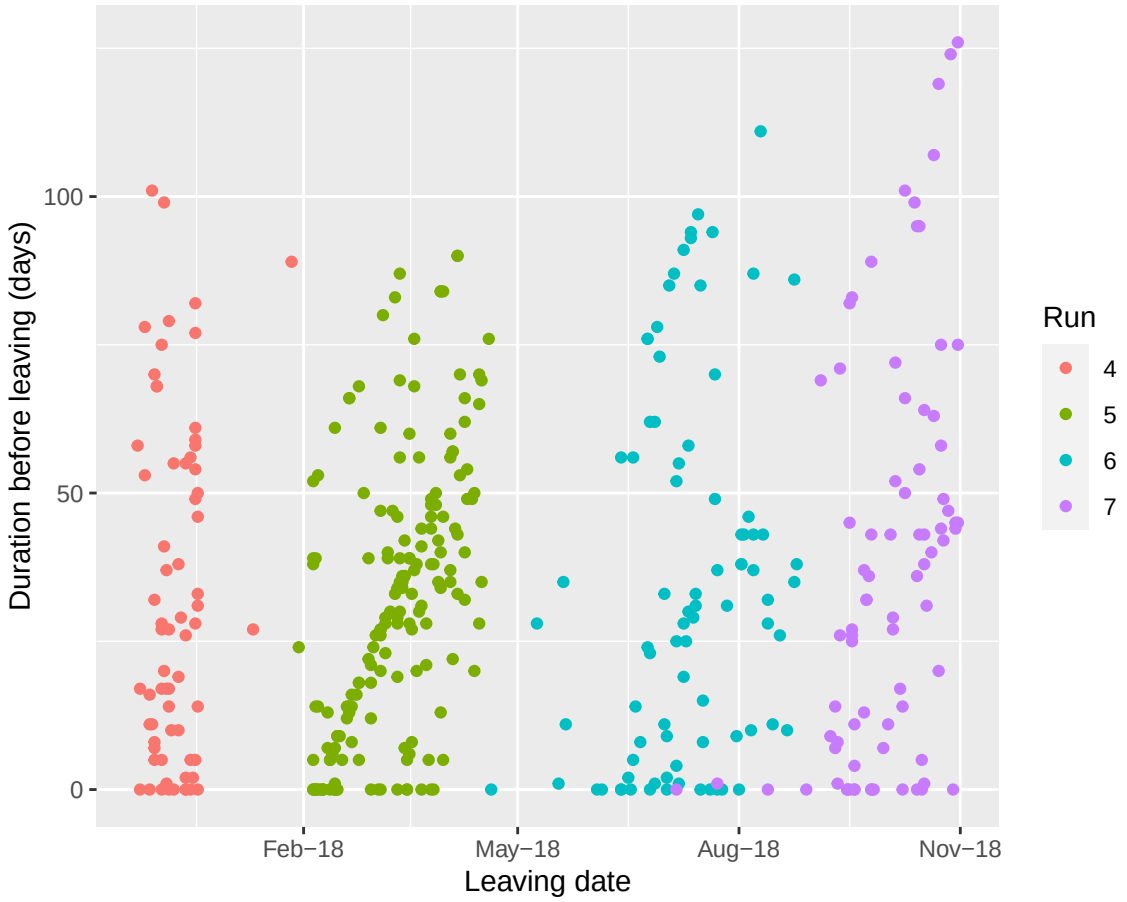
Lastly, we bind the data and quantify the students' leaving reason for easier graphing. We will then calculate the number of times each cause got picked. After presenting the graph, we will discuss it in the discussion area for a more detailed explanation.

```
#bind the data
merged_leaving <- do.call("rbind", list(leaving4, leaving5,
                                        leaving6, leaving7))

#quantify leaving reason for easier graphing
merged_leaving <- merged_leaving %>%
  group_by(leaving_reason) %>%
  mutate(reason_num = cur_group_id())
```

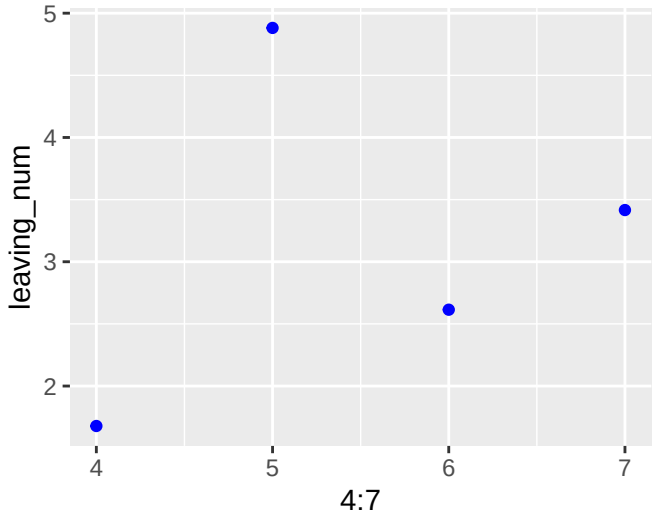
After the data preparation completes, we will begin to plot using the library of ggplot2. The third graph will demonstrate the duration from starting date until leaving date against the leaving date.

```
# Insert plot 3
graph3
```



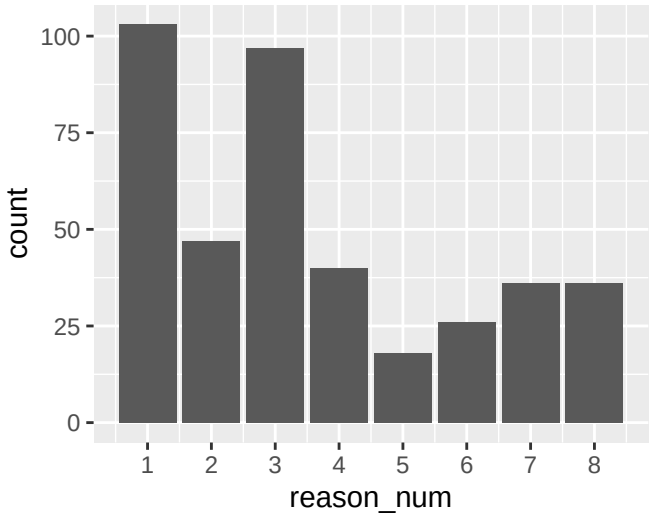
For the fourth graph, we will include the percentage of students who leave in the course in proportion to the overall enrollment for each run.

```
# Insert plot 4
graph4
```



The last graph will present the bar plot consisting number of entries for each reason.

```
# Insert plot 5
graph5
```



Discussion

In graph 3, we can see no distinct‘ patterns regarding either the leaving date or the duration from starting to leaving. Students leaving times are spread out evenly, just like the duration of quitting. Thus, we can not extract any valuable insight from this graph. For graph 4, the leaving rate of students (in proportion to the overall enrollments number) ranges from 2%-5% and is also relatively low. Comparing graphs 2 and 4, the data suggest that many students dropped the class and did not leave the course formally.

Graph five shows the bar plot consisting of the number of entries for each reason. Each number represents the reasons below

- I don’t have enough time (1)
- I prefer not to say (2)
- Other (3)
- The course required more time than I realised (4)
- The course was too easy (5)
- The course was too hard (6)
- The course wasn’t what I expected (7)
- The course won’t help me reach my goals (8)

We can see from the graph that the biggest reason for students leaving the course is reason one, which is “I don’t have enough time.” It then followed by reason three, “Other” and the other reason fairs about the same. The high count of reason one suggests that many students did not have enough time to learn and complete the course material. This interpretation could be

considered when designing the structure of future course runs. The teaching and quiz interval could be made more sparse, thus allowing more time flexibility for the students. Since many students answered "Other," it indicates that their reason for leaving is not stated in the options above. Thus, to enable a more directed feedback process in the future, more options should be listed, or students should have the chance to write their reason for leaving directly.

Question 3

What topic interests students the most?

For the third question, we will see which topic interests students the most through the video viewing status provided in the video-stats.csv. In particular, we will check variable viewed_one-hundred_percent and see which topic have the highest rating. We will begin by pre-processing the data first.

Additional notes

Configuration setting

Conclusion

AAA