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 also want to check what device students watch most of the video from. Thus, we want to check the correlation between the video views and the device form. After formulating the previous problems into a sentence, we come up with the following precise questions:

- Is the students in the course highly engaged?
- How many students and what causes students to leave the course?
- Which topic interests students the most?
- Is there any correlation between thevideo views and the device form?

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 try 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. We are also going to explain which data we are going to pick.
- 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 and num format.

ALSO STATE WHY YOU PICK THE FILE

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. For the graph result, we will include and discuss them in the *Modeling* and *Evaluation* phase.

Question 1

What is the participation rate of the students?

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
```

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())
```

Question 3

Which 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 total views and viewed_onehundred_percent. Then, we will discuss the results in the discussion section. We will begin by pre-processing the data first.

```
#read file video stats run 3 to 7 and store it in variable
files = list.files(path = "data/",
                   pattern="*video-stats.csv"
                   , full.names = T)

#the range start from 3 since run 1 and 2 does not provide
#the video stats
for (i in 3:7) {
  temp <- paste("video", i, sep = "")
  assign(temp, read.csv(files[i-2]))
}
```

After pre-processing phase, we can move on to processing the data. For each run, we will keep only the variables or columns related to our observations. Those variables are step position, title, total_views and viewed_onehundred_percent.

```
video3 <- video3[c(1:2, 4, 15)]
```

Next, we will index the data according to the run and bind them together to enable easier graphing process.

```
#index and bind the data to enable easier graphing
video3['run'] <- 3; video4['run'] <- 4;
video5['run'] <- 5; video6['run'] <- 6;
video7['run'] <- 7

merged_video <- do.call("rbind", list(video3, video4, video5,
                                      video6, video7))
```

In order to get accurate x-axis, we will convert the step_position from numeric format into string format.

```
merged_video$step_position <-
  as.character(merged_video$step_position)
```

Question 4

Is there any correlation between the number of views and the device form?

We want to check the correlation between the total number of views and the device form for problem four. We will only keep the related variables, which are video_duration, total_views, viewed_onehundred_percent, desktop_device_percentage, mobile_device_percentage, and tablet_device_percentage. Console and tv entries will be discarded since it only amounts to a little. After subsetting the data, we compute the correlation using the cor function. We will then store the result in a variable and discuss the result in the next phase.

```
cor_res <- cor(video_cor)
```

Modeling and Evaluation

This section combines two different phases for a more straightforward reading format. For the modeling part, we will graph our findings question by question, and for the evaluation, we will discuss any results we can derive from the graph.

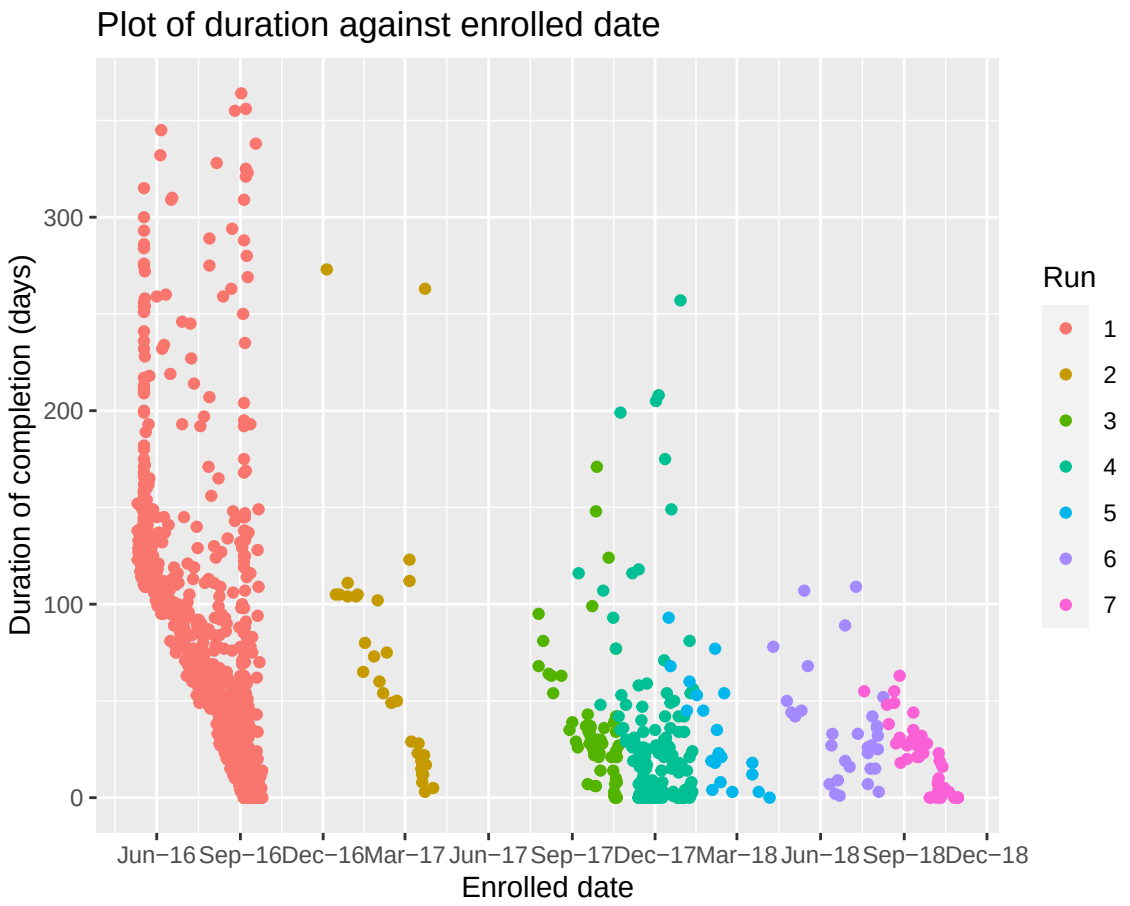
Question 1

What is the participation rate of the students?

For the first question, we want to analyse whether students are highly engaged in the course by checking the percentages of students fully participated and finished the material of the classes. Additionally, the duration of the completion for each student will also be checked. After finished preparing the data in the previous phase, we can now plot the graph.

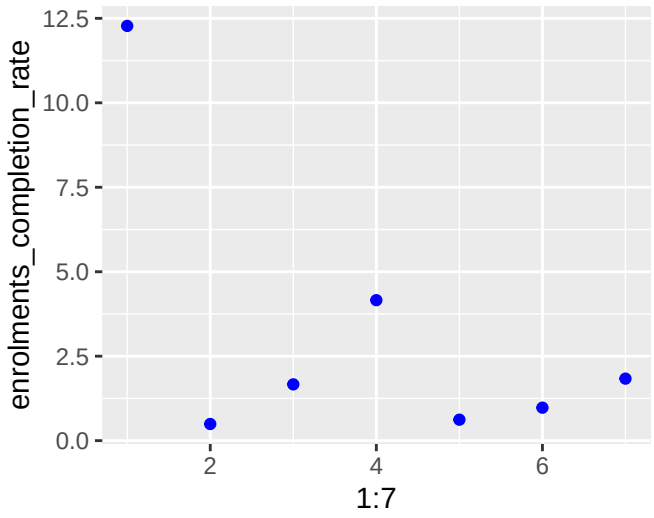
For the first graph, we will graph using ggplot2 to plot the duration of completion against the starting date for each student in each run.

```
# 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.

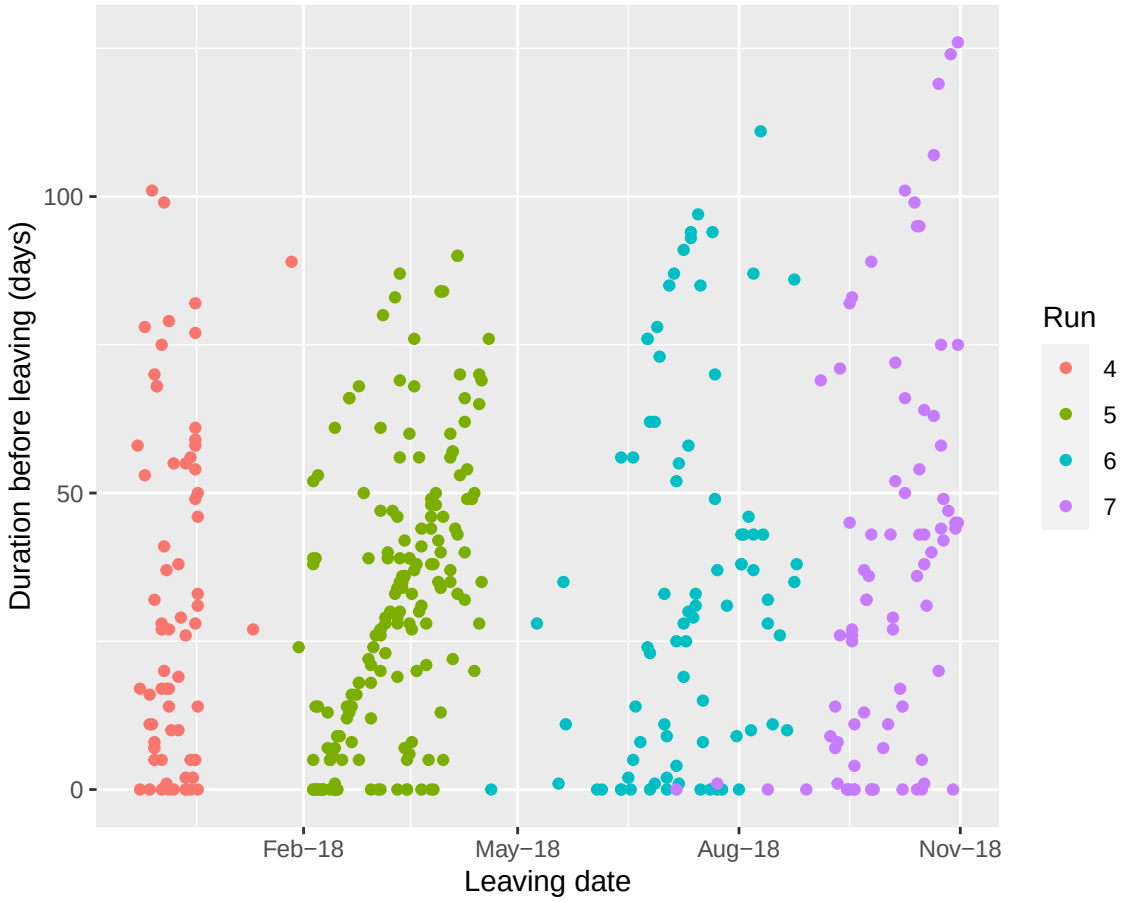
Question 2

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

For the second question, we want to evaluate number of students leaving the course, the time they quit, and the reason why. We will now plot the graph using library ggplot2 and the data prepared in the phase before.

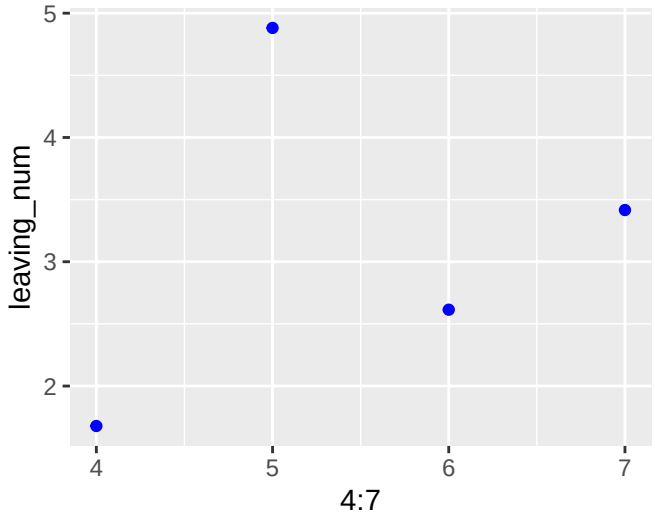
Graph 3 will demonstrate the duration from starting date until leaving date against the leaving date.

```
# Insert plot 3
graph3
```



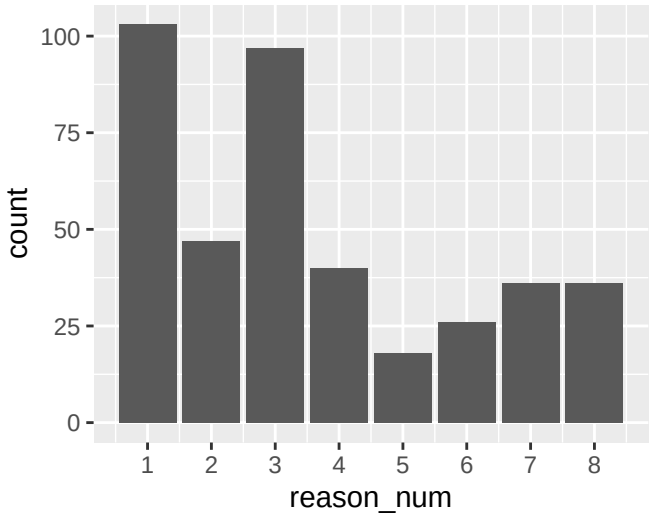
For graph 4, we will include the percentage of students who leave in the course in proportion to the overall enrollment for each run.

```
# Insert plot 4
graph4
```



The last graph (graph 5) 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 5 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.

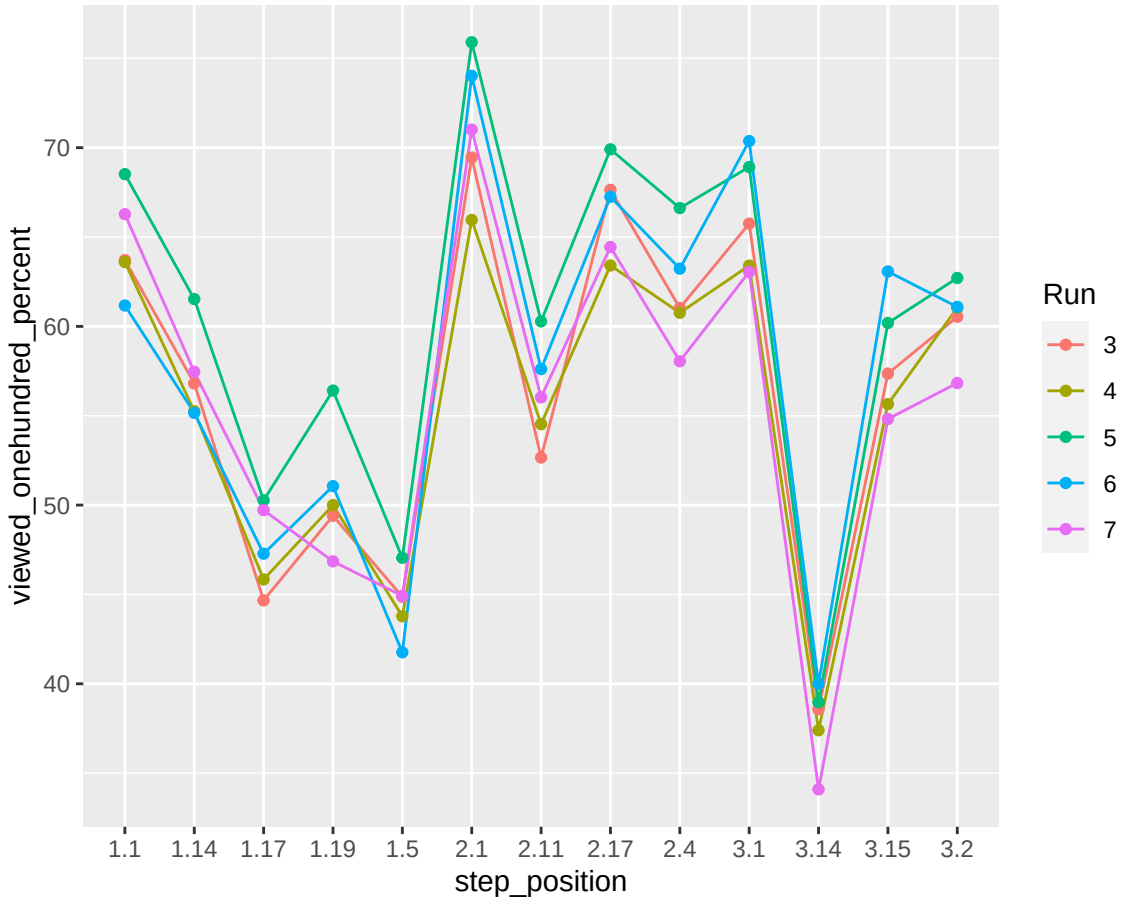
Question 3

Which topic interests students the most?

The video viewing status provided in the video-stats.csv will be used to determine which topic attracts students the most in the third question. We will look at the variables total_views and viewed onehundred percent in detail. We will use the data we collected in the previous phase to model the graph.

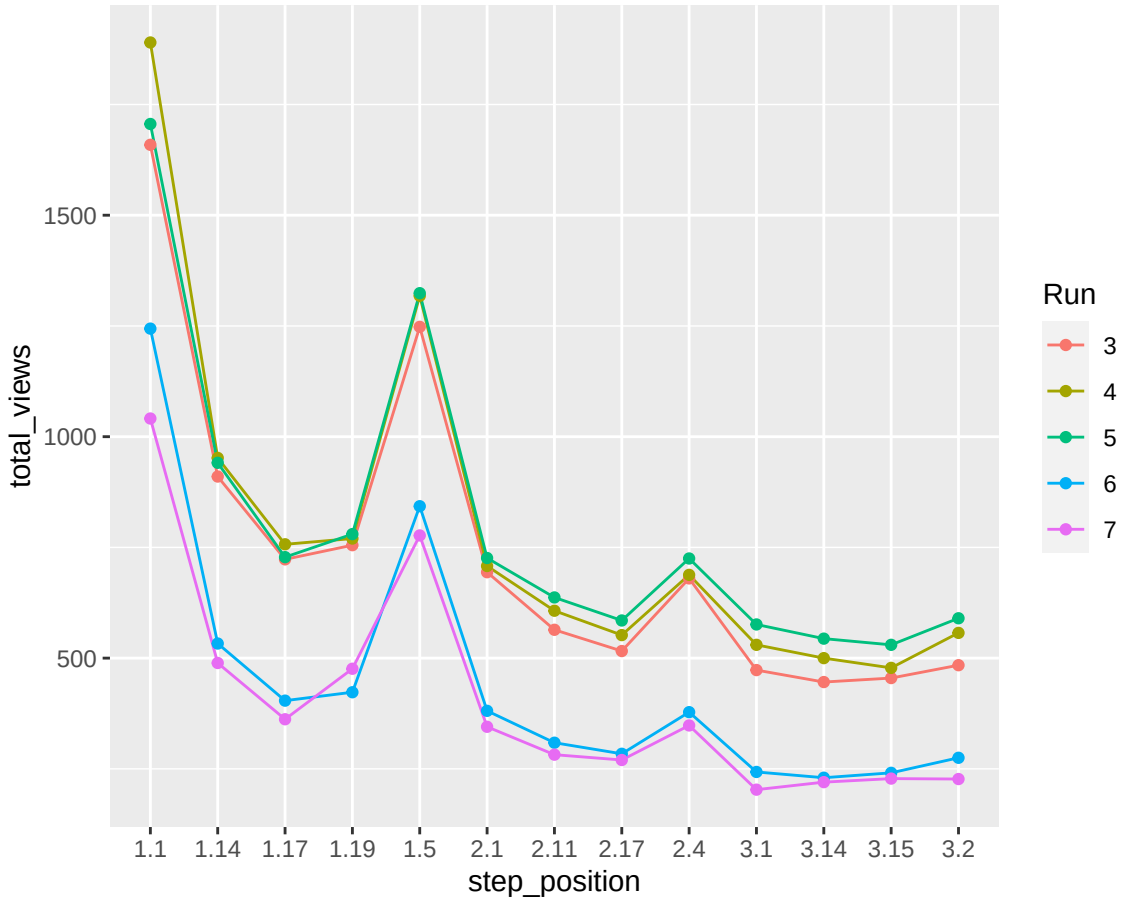
Graph 6 represents the variables viewed_onehundred_percent against the step_position. Each run defined by different colored line.

```
# Insert plot 6
graph6
```



Similar to graph 6, graph 7 represents the variables total_views against the step_position. Each run characterised by distinct colored line.

```
# Insert plot 7
graph7
```



Discussion

In graph 6, we can see that each run peaks at step 2.1, which corresponds to the topic of "Welcome to Week 2: payment security". This interpretation suggests that those topics are the most engaging ones for the student since many watched them until the end. For graph 7, the highest peak is at the first video, corresponding to the topic "Welcome to the course." The second peak is at step 1.5, corresponding to the video "Privacy online and offline." Overall, the trend of the graph is decreasing. This analysis suggests that most of the students are high-spirited at the beginning of the course, and their participation decreases as time goes by. The plot also indicates that the second peak topic, "Privacy online and offline," catches the students' attention the most. It is possible to elaborate more on the issues mentioned above, payment security, and online or offline privacy for future course runs.

Question 4

Is there any correlation between the video views and the device form?

To check whether there are correlation between the number of views and the device form, we already used the cor function in the previous phase. The results are as follow

```
cor_res
```

```
##                           video_duration total_views viewed_onehundred_percent
## video_duration                1.00000000 -0.07226244               -0.55223027
## total_views                  -0.07226244  1.00000000                0.01969757
## viewed_onehundred_percent    -0.55223027  0.01969757                1.00000000
## desktop_device_percentage     0.05909484 -0.18405570               -0.50509760
## mobile_device_percentage     -0.20594228  0.96364814                0.20339651
## tablet_device_percentage      0.15528062 -0.96390169                0.02344017
##                           desktop_device_percentage mobile_device_percentage
## video_duration                          0.05909484               -0.2059423
## total_views                            -0.18405570                0.9636481
## viewed_onehundred_percent              -0.50509760                0.2033965
## desktop_device_percentage               1.00000000               -0.3725834
## mobile_device_percentage               -0.37258342                1.0000000
## tablet_device_percentage               -0.02729193               -0.9076307
##                           tablet_device_percentage
## video_duration                          0.15528062
## total_views                            -0.96390169
## viewed_onehundred_percent               0.02344017
## desktop_device_percentage              -0.02729193
## mobile_device_percentage               -0.90763066
## tablet_device_percentage                1.00000000
```

Discussion

From the correlation result, we can see a strong positive linear relationship between the total_-views and mobile_device_percentage. This result suggests that most students watch the lecture video from their mobile phones. The correlation matrix also shows a positive relationship between viewed_onehundred_percent with mobile_device_percentage and tablet_device_percentage. This finding indicates that students watching from their mobile or tablet devices are more likely to finish their video than the other students watching from the desktop. Other than that, the correlation matrix does not show any other distinct patterns.

Deployment

For the deployment part, the user can access the whole process on the following Github repository. Additionally, this written final project will also be one of the deployment forms since the user will understand and gain insight through reading this.

Conclusion

In conclusion, we have completed the exploration process of a data science project while adhering to the CRISP-DM methodology. We managed to answer the questions imposed in the *Business Understanding* part with coherent solution explained in the several phase after that. Resources introduced in the course, such as library ggplot2 and dplyr, are also used throughout the process.

kurangin number of codes (email joe)
final check everything