CRISP-DM Model on MOOC Dataset

Subin Jung

2022-11-02

Business Understanding

Business Objectives

Learning Analytics is becoming one of the important subjects for Data Science. It is often recognised as the measurement of collection, analysis and reporting the data of learners and their surroundings. Especially, the main purpose of learning analytics focuses on delivering shared insights and finding efficient method of measuring engagement of learners.

The massive open online course (MOOC) dataset, called "Cyber Security: Safety At Home, Online, and in Life", is collected from its online method provider FutureLearn. It has been collected by 7 runs of online course and contains the learners' progress and their profiles.

The main business objectives is understanding who the main client base is and comprehending their needs in order to enabling to provide learners more essential learning materials. Examples of related business questions could be "What is the background/job of learners and does it affect to the engagement of course?" and "What is the average time to complete the course for learners and why?"

Situation Assessments

Data Available MOOC dataset includes lots of information which consists of one *5 step names overview.html* file and 7 iterations of following data:

- archetype-survey-responses.csv
- enrolments.csv
- leaving-survey-responses.csv
- question-response.csv
- step-activity.csv
- weekly-sentiment-survey-responses.csv
- Course Overview - FutureLearn Course Creator.pdf

Hence, this study will selectively investigate on the *enrolments.csv* and *step-activity.csv* data to focus on the learner's enrolment background and the time intervals of studying progress.

Requirements, Assumptions and Constraints This project is free from any security and legal restrictions. Also, it is planned to finalise within three weeks by one data analyst.

Moreover, since the dataset have been given already, no economic factors will be produced. In terms of data, it is assumed that the data quality is accurate since it is collected by registered learners limitedly. Finally, the results will be reported by the original template code, analysis report and the presentation.

The data is accessible easily by download, however, there would be no legal constraints since all the individual's data is protected by anonymous IDs instead of learner's name.

Risks and Contingencies There could be a risk which can lead to misinterpretation if there are missing data more than saved one. In addition, since the project is performed by one data analyst, the data of completion could be possibly delayed. Furthermore, since the data is collected from a realistic situation, there might be no outstanding discoveries.

Data Mining Goals

Goals

The business goal is decreasing the unenroll ratio of learners by discovering the reason of unenrolling. To achieve this, the goals for the study are specified as follows:

- Classify the background information - for example, occupation, highest education level and nationality - from enrolment data to discern the main target learners.
- Compare the ratio of enrolments/unenrolments and identify the learners' background who unenrolled.
- Calculate the average completion time for each step from step activity data to find the relation between course material and unenrolments.

Data Understanding

Initial Data Collection & Description

Promising data

- **Enrolments** The enrolments dataset includes learner's information about the date of enrolment/unenrolment and their background with learners' IDs. Especially, the enrolment/unenrolment date seems useful to identify the ratio of unenrolled learners and the background information could help to discern the main target clients.

  However, the data quality can be improved by removing missing values since there are lots of unknown values in several background columns. Furthermore, although the dataset has 7 runs, it is decided to use only the first run as the first-year data already contains more than 100,000 rows, which is enough to generalise the conclusion.

- **Step Activity** The step activity dataset contains the module number consists of two numeric value - week number and step number - and the first visiting time/last completion time with learners' IDs. In other words, it is showing the each of step's completion time for every learner. This might be helpful to compare the time taken among the steps by calculating the average time so that it is possible to identify which steps take the most and the least time to complete. Eventually, it is expected to recognise between the proper and necessary steps and those are not.

  Compared to enrolments data, the quality of step activity data is high since there are no missing values without some of cells from the last completion time columns. Nevertheless, those missing rows give important information, which can be consider ed as the learner who did not complete the step. In addition, in the same context as the enrolments data, the data has more than 100,000 rows so only the first round is enough amount.

Irrelevant data

Archetype Survey Responses, Leaving Survey Responses and Weekly Sentiment Survey Responses dataset are revealed that all the data are missing so it is impossible to analyse for project goals. Also, although there are data in the Question Response dataset, it is about the learners' answer of quizzes so it is hard to detect the relevance to the project goal.

Data Exploration

In enrolments dataset, there are lots of missing values in employment status, highest education level and country attributes. This can prevent from meaningful exploration of learner's background so it is suggested to remove the rows which involves any missing cells.

However, step activity dataset is available to explore as it has fewer missing cells. Using the time stamp of first visiting and last completing data can explain the average time spent on each step activity. Furthermore, by comparing the average time spent, it is feasible to tell what kind of supplements are suitable for learners.

Data Quality

As clarified in the previous paragraph, the enrolments dataset has some missing values, nonetheless, these will be deleted before the examination for higher data quality. Also, even though the step activity data has few of missing values in the last completing time attributes, it can be ignored for deletion and regarded as non-complete learners.

Data Preparation

Selecting Data

In earlier stages of the data mining procedure, many choices about which data and attributes to choose have already been set. To be specific, among the background information from enrolments data, highest education level, employment status and country attributes will be used and the two given time attributes is selected from step activity dataset.

Cleaning Data

To resolve the issues raised in the data quality report, data cleaning is implemented. The missing data are withdrawn from entire enrolments dataset. On the other hand, the step activity dataset is remained as original since the data quality is adequate to be inspected.

Constructing Data

Several new features and records are derived to append in the dataset.

- The "enroll_1" and "activity_1" records are the main dataset which only drop missing values from originals.
- The "Unenroll" record is extracted from enroll_1 but only contains unenrolled learners' information.
- The "complete time taken" attribute is calculated by differentiation between the first visiting time and the last completing time for step activity.
- The "completed_act" record is extracted from activity_1 but only have completed activities.
- The "longest top 10" record has the identical attributes as step activity record but only composed of the activities within 10th longest completion time.
- The "shortest top 10" record is only made up of the activities with the 10th shortest completion times, but it has the same attributes as the step activity record.
- The "mean of step" record is only consist of two attributes, the step activity numbers and the corresponding average time calculated.

Formatting Data

- Those two time attributes from step activity dataset are converted from character type into date type to calculate the time intervals. At last, both are set as numeric value to plot a graph.

- The step activity number is converted into character type from integer type to plot a box plot by class.

Modelling

Techniques

- Grouped bar chart of employment status by highest education level
- Bar chart of employment area and country (Top 10)
- Ratio of enroll/unenroll
- Pie chart of employment status among unenrolled learners
- Bar chart of employment area and country among unenrolled learners
- Ratio of complete/non-complete
- Distribution of completing time of activity in total
- Box plot of average time taken for step - top 10 longest and shortest time

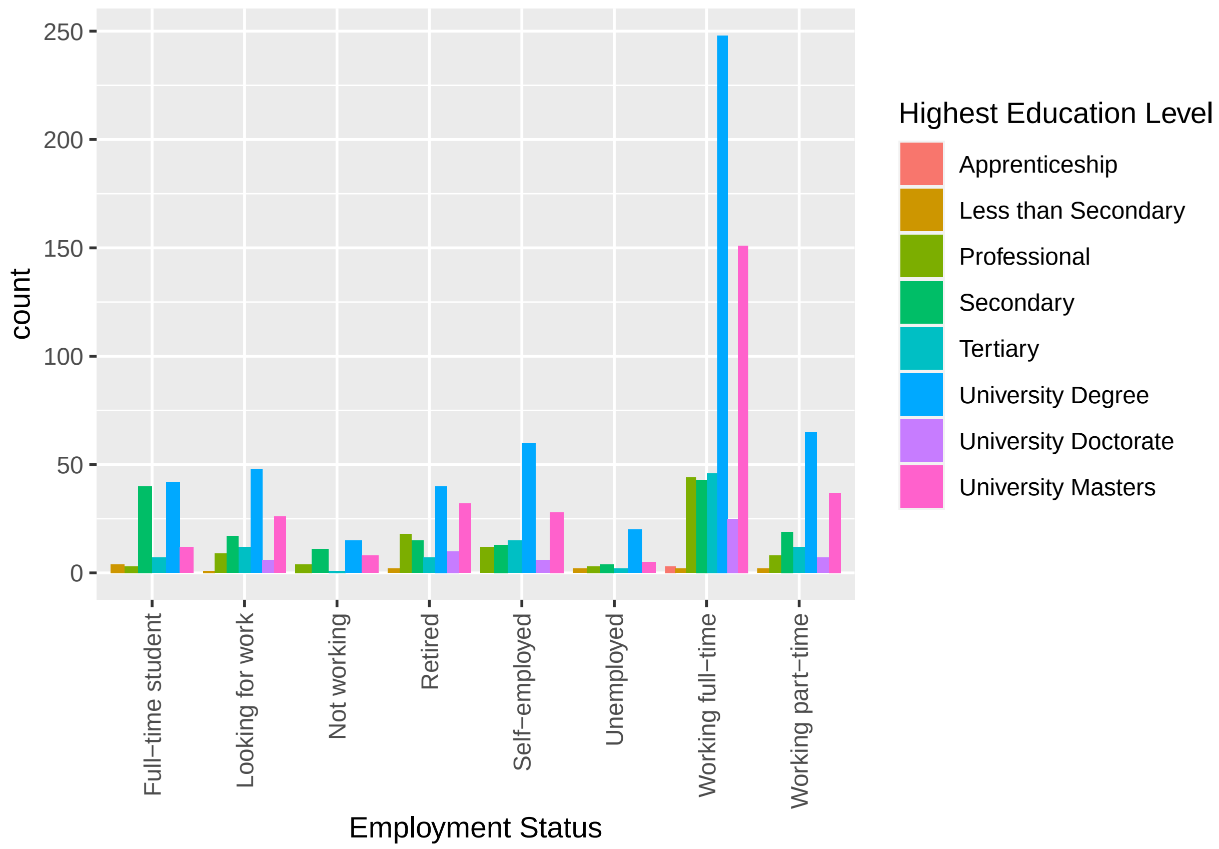
Models

1st Cycle: Which background of people register and unenroll the most? (Find our main learner target based on their background)

To identify the main target learner who register at first place, the investigation for their background - such as employment status, highest education level, employment area, country - is implemented.

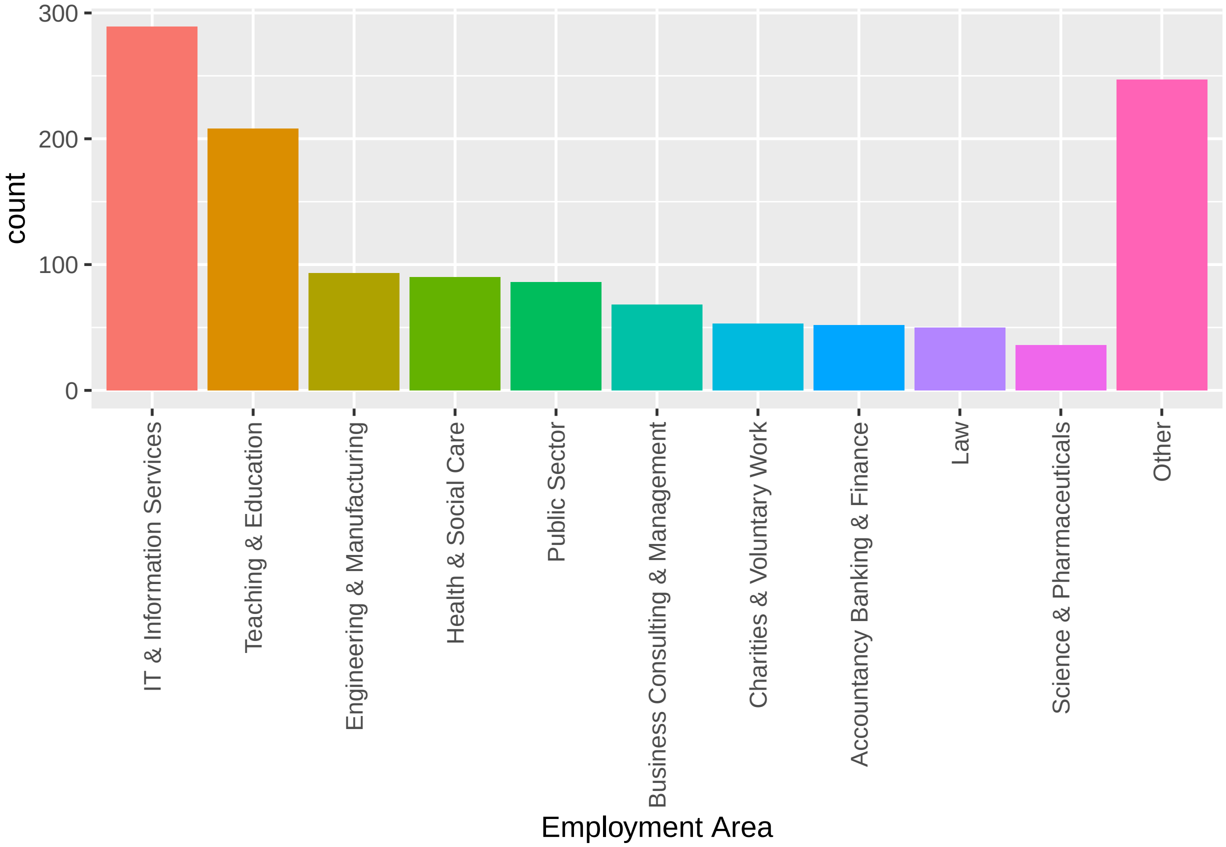
Distribution of employment status by highest education level As the following grouped bar chart describes, it is clear that most of users are full-time worker and have university degree or master's degree as their highest education level. Also, regardless of the highest education level, full-time worker has the most amount compared to other employment status. Therefore, it can tell that the main client base who register is full time workers.

Interestingly, although it is an education material, the ratio of full-time student is quite low. It can be interpreted that, rather than full-time students, people who graduate highest education seeks for more academic course during their working life.

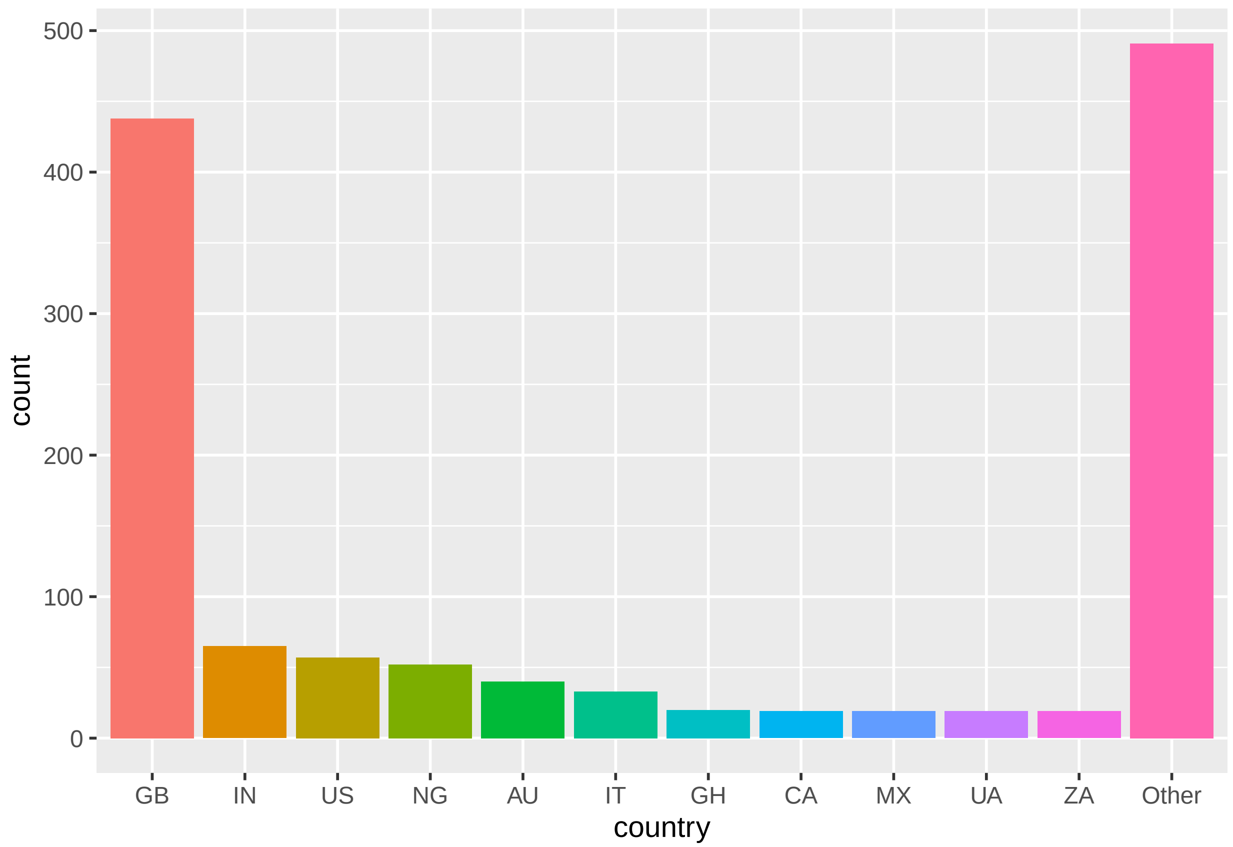


Distribution of employment area & country (Top 10) Since there are over 20 categories both in employment area and country, the data is segmented to focus on the highest 10 values.

As the following bar chart shows, those who work in IT & Information Services area is almost 300 people as the most and the second highest is around 200 people who works in Teaching & Education sector. According to this information, the purpose of register the course could be not only learning the skills for themselves to use but also for education purpose to deliver the knowledge to other people.



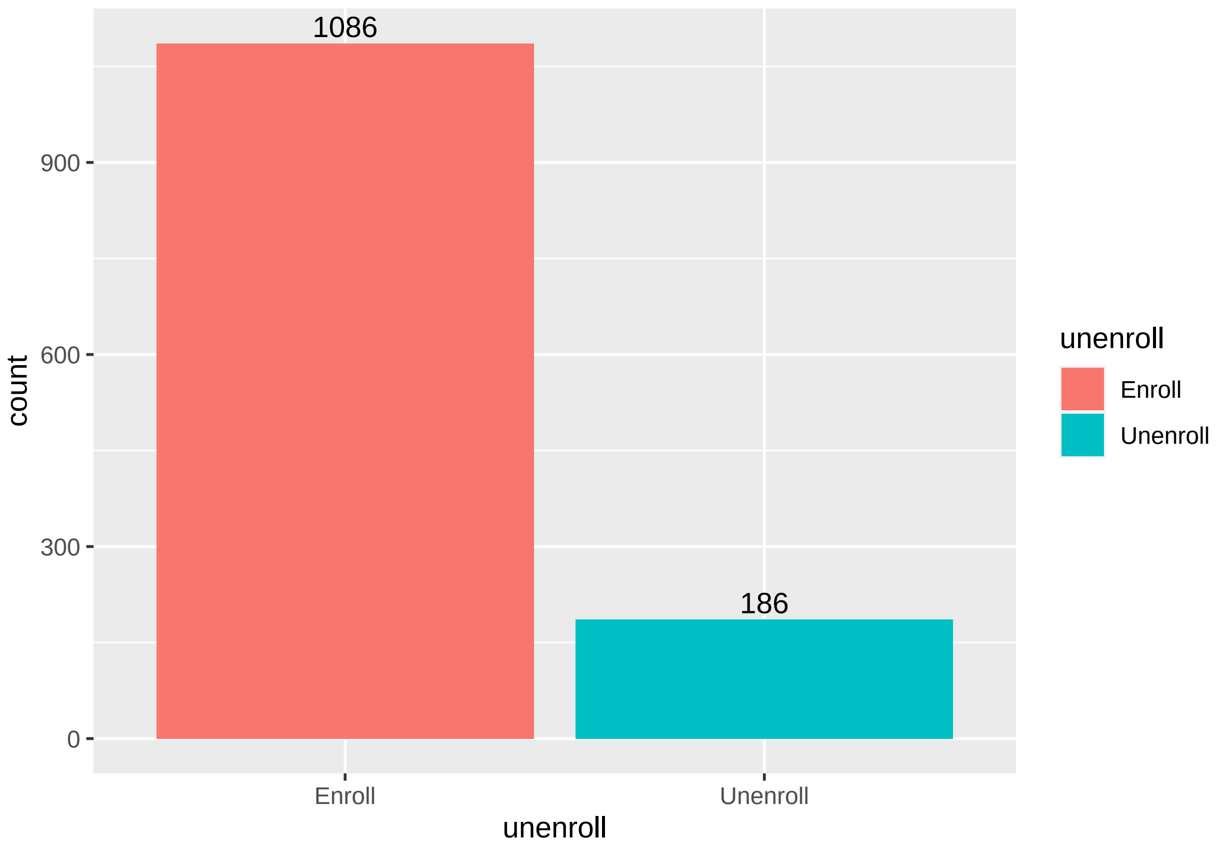
Additionally, based on the bar chart below, people who are from United Kingdom enrolled overwhelmingly showing that more than 400 people. In contrast, from the second highest country, they only registered less than 100 respectively. In terms of this result, our target client is distributed in United Kingdom in general.



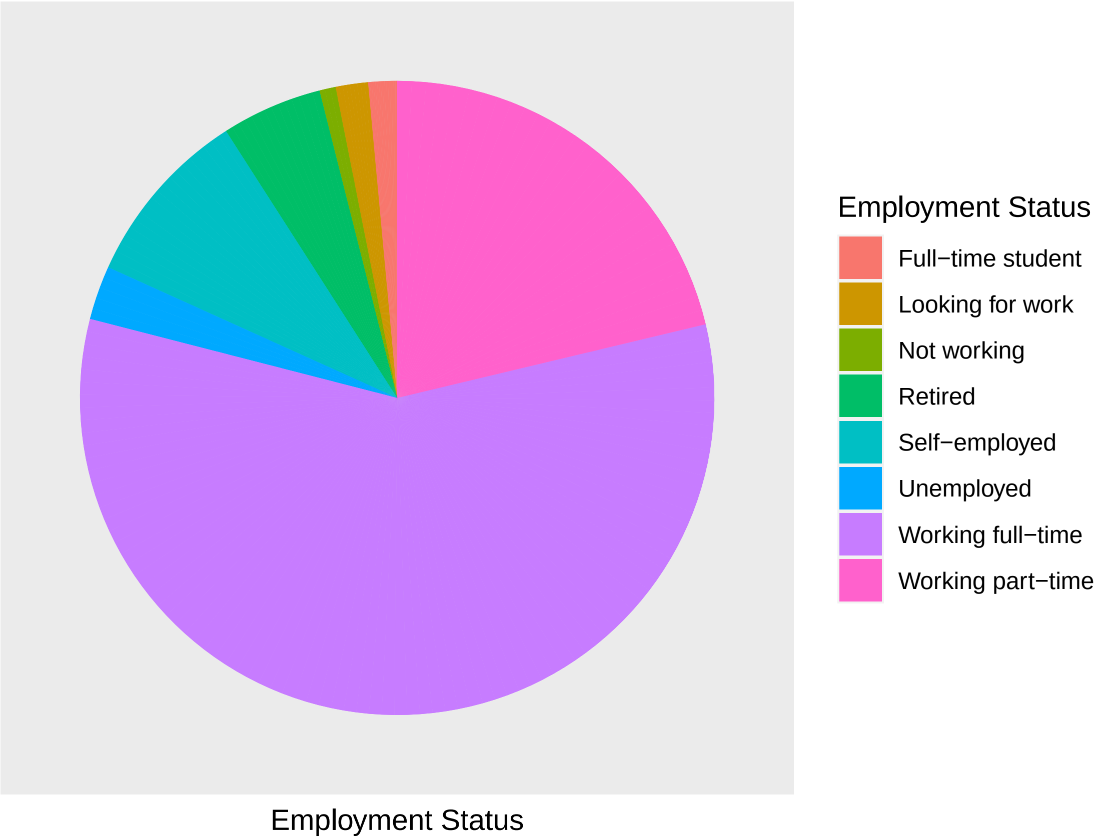
Ratio of enroll/unenroll Even though studies for the main client base is executed above, it is crucial to understand the unenrolled users to prevent the loss in the future.

According to the chart below, the number of total enrolments is 1086 and the unenrolled users among them is 186 which is 17% of registered users.

```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(count)` instead.
```

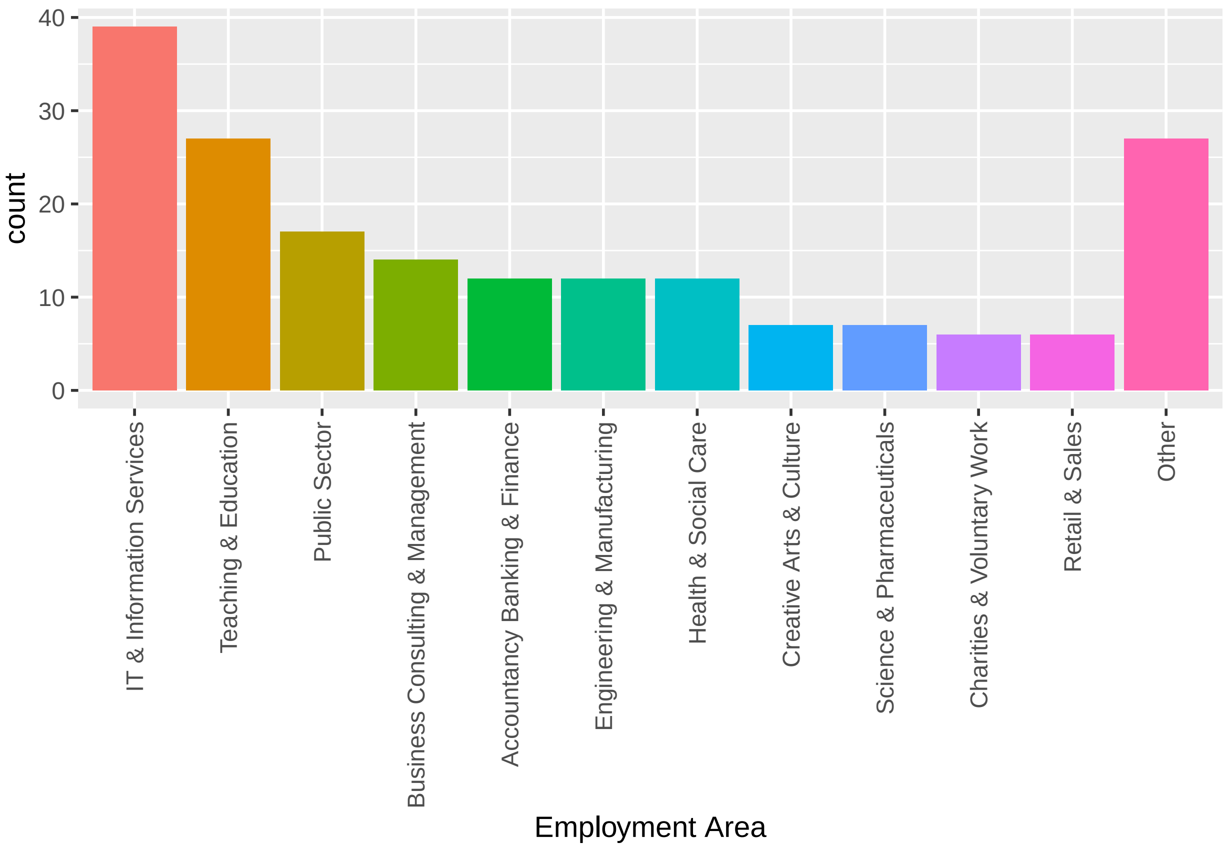


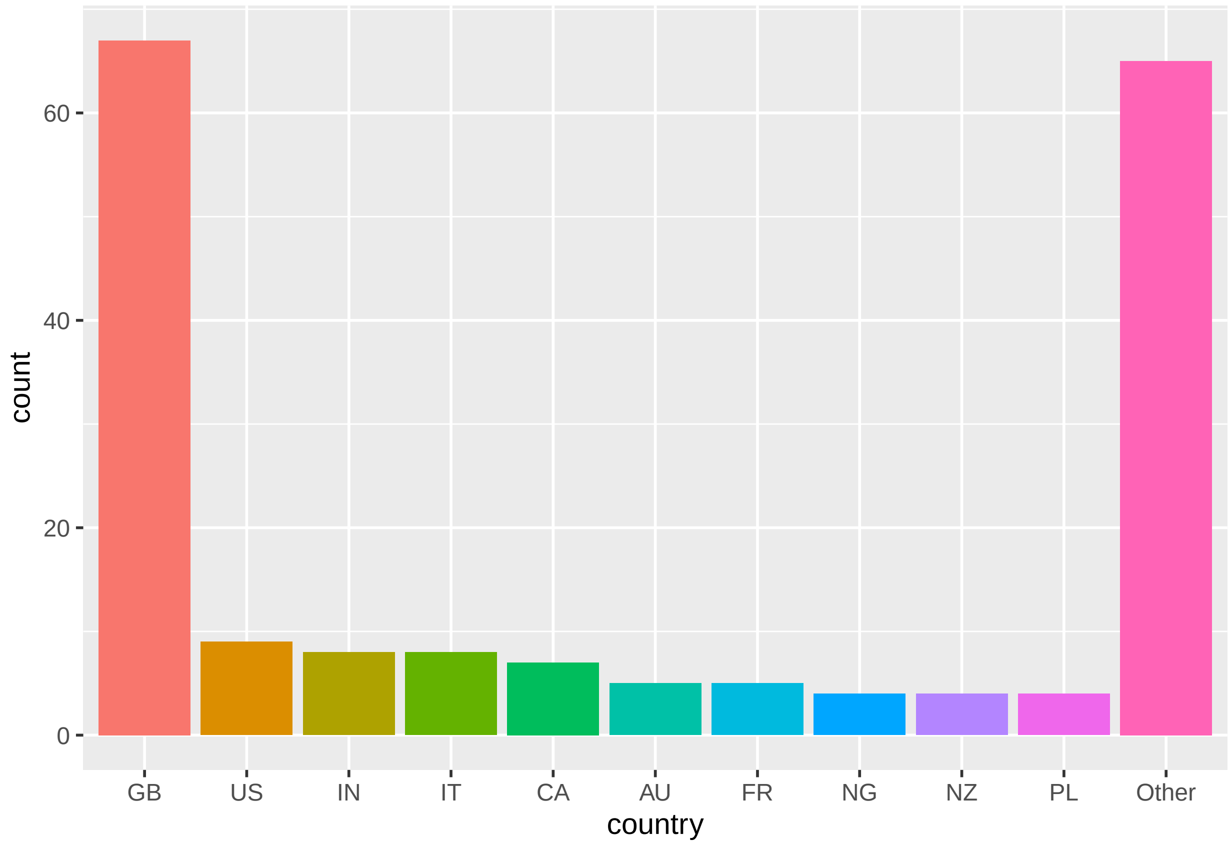
Pie chart of employment status among unenrolled learners As one of the unenrolled learner's background information, employment status is inspected with the pie chart as follows. Similar to the total registered users' data, full-time worker is the most common employment status than others. In other words, employment status might not influence user's unenrolling.



Bar chart of employment area and country among unenrolled learners In the same vein, as following two chart display, the employment area and the country rate also similar as the result of total learner's data. Even if some of the categories within 10 differ, the first rank is the same in both charts.

In conclusion, it appears that there was no difference between enrolled and unenrolled learners in terms of their profiles, thus, the background is not a factor of unenrolment for learners.

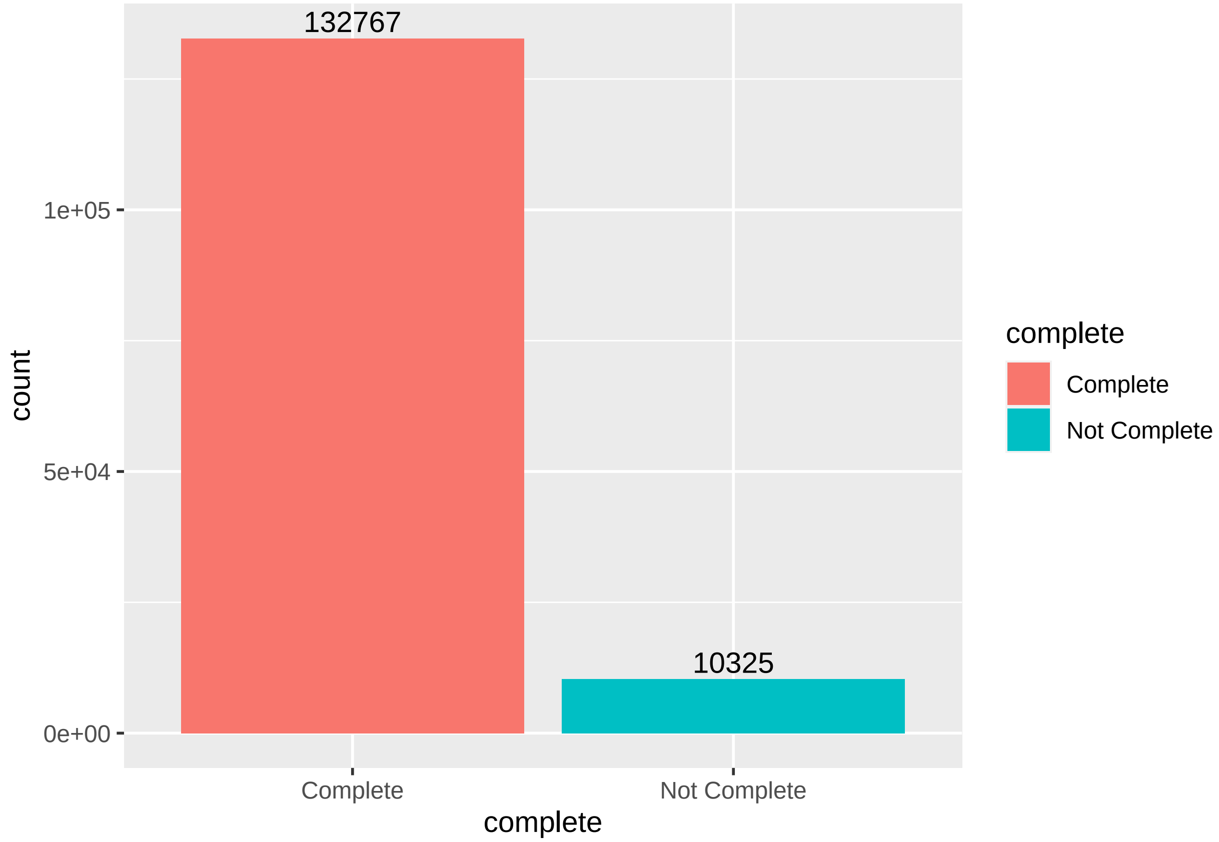

count
40
30
20
10
0
IT & Information Services
Teaching & Education
Public Sector
Business Consulting & Management
Accountancy Banking & Finance
Engineering & Manufacturing
Health & Social Care
Creative Arts & Culture
Science & Pharmaceuticals
Charities & Voluntary Work
Retail & Sales
Other
Employment Area



2nd Cycle: Does the course material influence to ending up unenrolment?

In the previous cycle, it turns out that background has no relation with the reason of unenrolling. To find out the other source of impact, the analysis for course material is performed.

Ratio of complete/non-complete Corresponding to the figure below, the majority of activities are completed and only 7% of course is not finished.

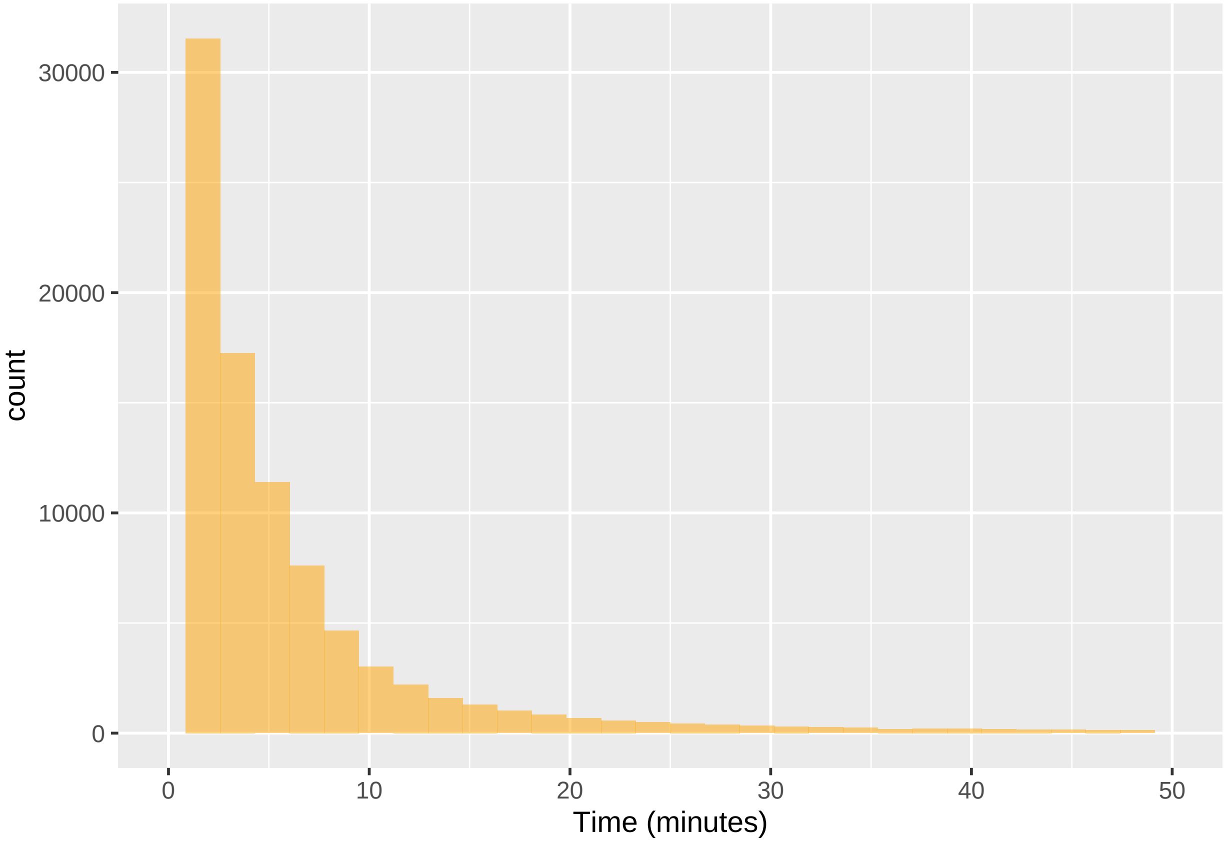


Distribution of completing time of activity in total Besides, as the histogram of completing time below describes, over 50000 activities are completed in 10 minutes. However, it is difficult to generalise the time with all activities since the material of the step is all different such as video, article, exercise, quiz and so on.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 26212 rows containing non-finite values (`stat_bin()`).
```

```
## Warning: Removed 2 rows containing missing values (`geom_bar()`).
```

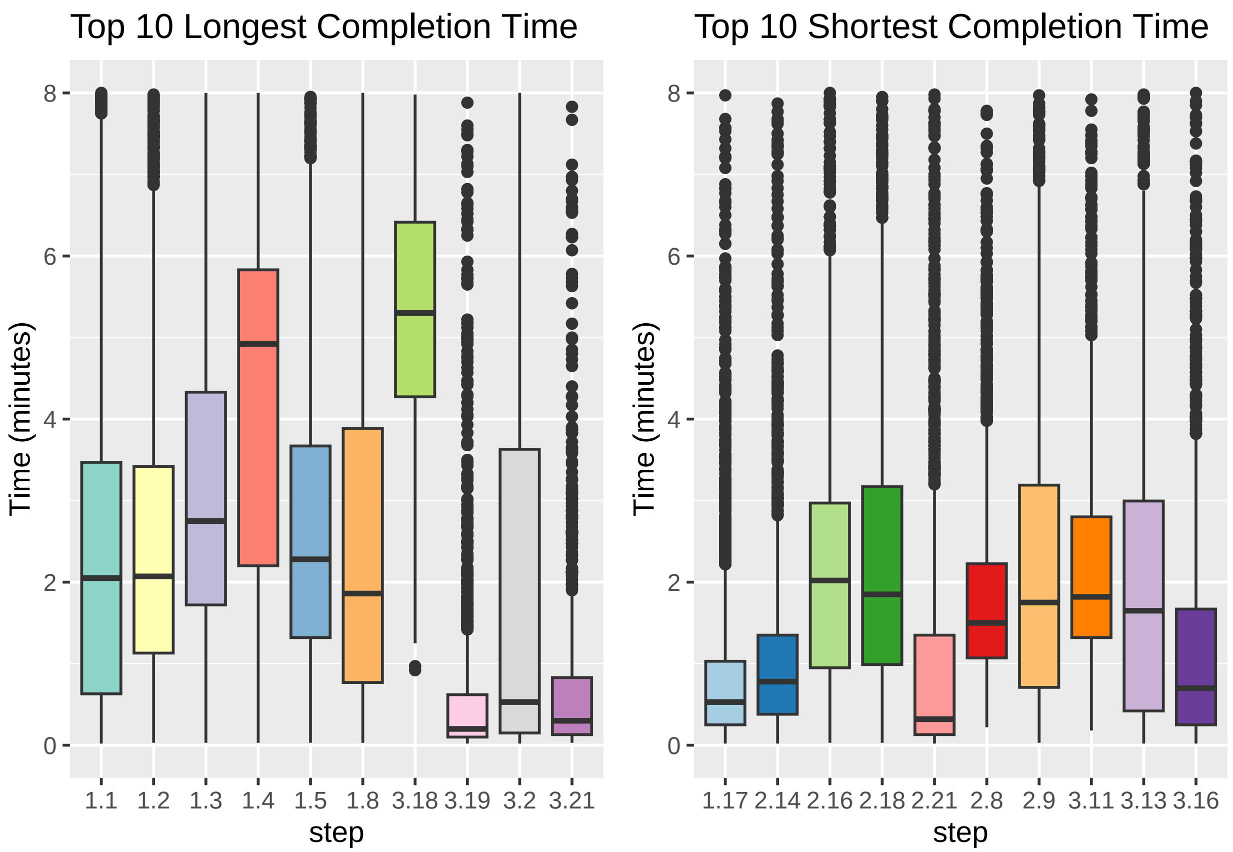


Box plot of average time taken for step - top 10 longest and shortest time Hence, for the analysis in detail, the list of the longest and shortest time of step is segmented by their average of all learners' completion time. From the first box plot below, the longest completion time taken among 10 steps is step 3.18 which is a test material. However, the second longest average time taken is step 1.4 which is a video of 4 minutes long. Besides those two, there are several other materials such as 5 articles, 1 exercise and 1 discussion.

On the contrast, according to the second box plot, the first shortest is step 2.21, discussion for summary and the second shortest is step 1.17, article and others are 4 articles, 1 video, 3 discussions and 3 quizzes. Compared to the longest top 10, it has more videos, discussions and quizzes. Nevertheless, the longest average time is still less than 6 minutes so it is hard to say the difficulties or completion time is affected on the reason of unenrolments.

```
## Warning: Removed 15807 rows containing non-finite values (`stat_boxplot()`).
```

```
## Warning: Removed 3614 rows containing non-finite values (`stat_boxplot()`).
```



Evaluation

Assessments

The business goal explained previously is decreasing the unenroll ratio of learners by discovering the reason of unenrolling. By performing 2 cycles of CRISP-DM, it is successfully examining the background of registered learners and unenroled learners. Nonetheless, the surroundings of learner do not have a relation with their reason of unenrolment since the background between enrolled and unenrolled users has no outstanding differences.

Furthermore, the study of analysing the step activities' average time succeed to discover the material difference between top 10 longest and shortest steps. However, the completion time is all similar as less than 6 minutes so it cannot be concluded that the course material influences on unenrolments as well.

Future works

Even though the given dataset has the 7 runs, only the first-year data was analysed for this report due to the limitation of time. Especially, in the data understanding paragraph, it is judged that the amount of data is enough to generalise the conclusion but after preprocessing the data amount has decreased because of the poor data quality. Consequently, it is suggested that analysing the data pf other 6 runs for improving the result accuracy.

Deployment

The deployment for this paper will be in presentation form with highlighting some of important points. The presentation will mention the business goal to help the understanding of two questions generated. Also, the data selected to identify those two cycle's questions - 1st cycle: Which background of people register and unenroll the most?, 2nd cycle: Does the course material influence to ending up unenrolment?, data: First year of Enrolments and Step Activity - will be illustrated. For the analysis, it will summarise the data preprocessing, modelling and evaluation part and the steps taken in each sector. Finally, two graph - Distribution of employment status by highest education level and Box plot of average time taken for step: top 10 longest and shortest - will be plotted and explain the results that can be interesting to stakeholders.