Data Management and Exploratory Data Analysis : Learning Analytics

Kaustubh Kulkarni | 230195431

2023-11-17

Introduction

This report is about the Learning Analytics on the MOOC(Massive Open Online Course) entitled "Cyber Security: Safety At Home, Online, and in Life" made by Newcastle University and made available to the public by the online skills provider FutureLearn. In this report, I will analyse the FutureLearn dataset and provide some significant insights to the provider FutureLearn. To carry out this investigation, I will describe two CRISP-DM cycles as follows:

CYCLE 1

1. Business Understanding:

The first step of the CRISP - DM cycle is to describe the business understanding. In this, I will define the objectives and the success criteria of this investigation. My task in this cycle will be to clear the objectives by trying to answer the business question and meeting the success criteria by implementing required techniques.

1.1 Business Objectives :

- The main objective of this report is to give an overview of the audience i.e. the learners of this course to the provider.
- To help the provider understand the background and demographic profiling of learners across all runs of the course.
- To assess in deciding which demographic area of the learners is to be focused more and what kind of audience must be targeted.
- Understand the trends and popularity of the course across all runs.

1.2 Success Criteria :

- The data used should be accurate and from the right source to produce relevant results.
- The implementation and insights should be able to complete the business objectives.
- The analysis should derive outputs that will help provide solutions to the problems and questions raised.
- The documentation must be clean and easily understandable to the providers in order to take actions needed based on our suggestions.

1.3 Data Mining Goals :

- The data must be extracted from original dataset and in the right necessary formats.
- The data should be cleaned and preprocessed for analysis.
- Demographic analysis should be done on the data to complete our objective.
- Analysis done on the data should provide right solutions to our problems with the help of right choice of visualizations.

1.4 Research Question :

With respect to the objectives, this report tries to answer the following question:

What are the general demographics of learners participating in the MOOC "Cyber Security : Safety At Home, Online and in Life" offered by Newcastle University through FutureLearn?"

2. Data Understanding :

This is the second phase of our cycle - Data Understanding - which in it's name itself states that it consists all brief information of the data like, from where data is collected, for what it is meant and the reason to be chosen. In this phase I will describe the data, it's sources and also the quality of data provided to me.

2.1 Data Collection :

The data is collected from the dataset provided by FutureLearn of all runs of the course. There were separate set of data files for each run and in different formats. There were .csv files which had all the stats and data whereas, there were pdf files of each run which consists of how webpage of course of each run looks like. Here I am focused on the .csv files which will be relevant for my objectives. For our objective, I have chosen the enrollment data which is stored in "enrollments.csv" file of each run. The data from files for each run is imported into dataframes and integrated for further use.

2.2 Data Quality :

As discussed earlier, I have chosen the enrollment data for our objective, and when I checked the files the quality of data was inconsistent. The information below will show the description and the quality of each column from the enrollment data file. This is an overall review for enrollment files of all runs:

- learner_id - This is the main id of learners and this column is clean and does not have any missing or unknown data.
- enrolled_at - This column states the time stamp the learners enrolled for the course. The data type of this column should be Date rather than General(no specific format).
- unenrolled_at - This column states the time stamp the learners left the course. Like above, this column's data type should also be Date. This column has many missing values.
- role - The role of each participant is given in this column and has only two values - learner and organization admin. This column is clean without any missing or non-format values.
- fully_participated_at - The time stamp when the learner fully participated in the course. This column is not in Date format and has missing values.
- purchased_statement_at - The time stamp at which the learner purchased the statement of the course. This column is not in Date format and has missing values.
- gender - This column states the gender of the learner. The data in this column has many unknown values. The data is not consistent.
- country - This column states country of the learner. The data in this column has many unknown values. The data is not consistent.

- age_range - The range of age in which the learner lies. The data in this column has many unknown values. The data is not consistent.
- highest_education_level - This column states the education level of the learner. The data in this column has many unknown values. The data is not consistent.
- employment_status - The employment status of the learner. The data in this column has many unknown values. The data is not consistent.
- employment_area - The area of employment of the learner. The data in this column has many unknown values. The data is not consistent.
- detected_country - The country detected from the system of learner. This data has many missing values and is inconsistent.

3. Data Preparation :

This phase focuses on the data preprocessing techniques starting from the selection of right filtered data for our research. I will perform steps like collecting, integrating, cleaning and preprocessing the data in this phase. All the major preprocessing of the data is performed in the preprocessing script in the 'munge' folder of the project directory. This phase provides the cleaned and filtered data to create different visualizations for our EDA which will be discussed in next phase.

3.1 Data Selection :

First as discussed in data collection, I imported the data into dataframes and integrated them together by adding run_number column to each row which stated from which run the data of learner is. After merging the data, I calculated the number of learners in each run to kickstart a trend. But, the data is not clean yet as it has many irrelevant columns and missing and unknown values.

```
## # A tibble: 7 x 2
##   run_number num_learners
##        <int>        <int>
## 1          1        14394
## 2          2         6488
## 3          3         3361
## 4          4         3992
## 5          5         3544
## 6          6         3175
## 7          7         2342
```

3.2 Data Cleaning :

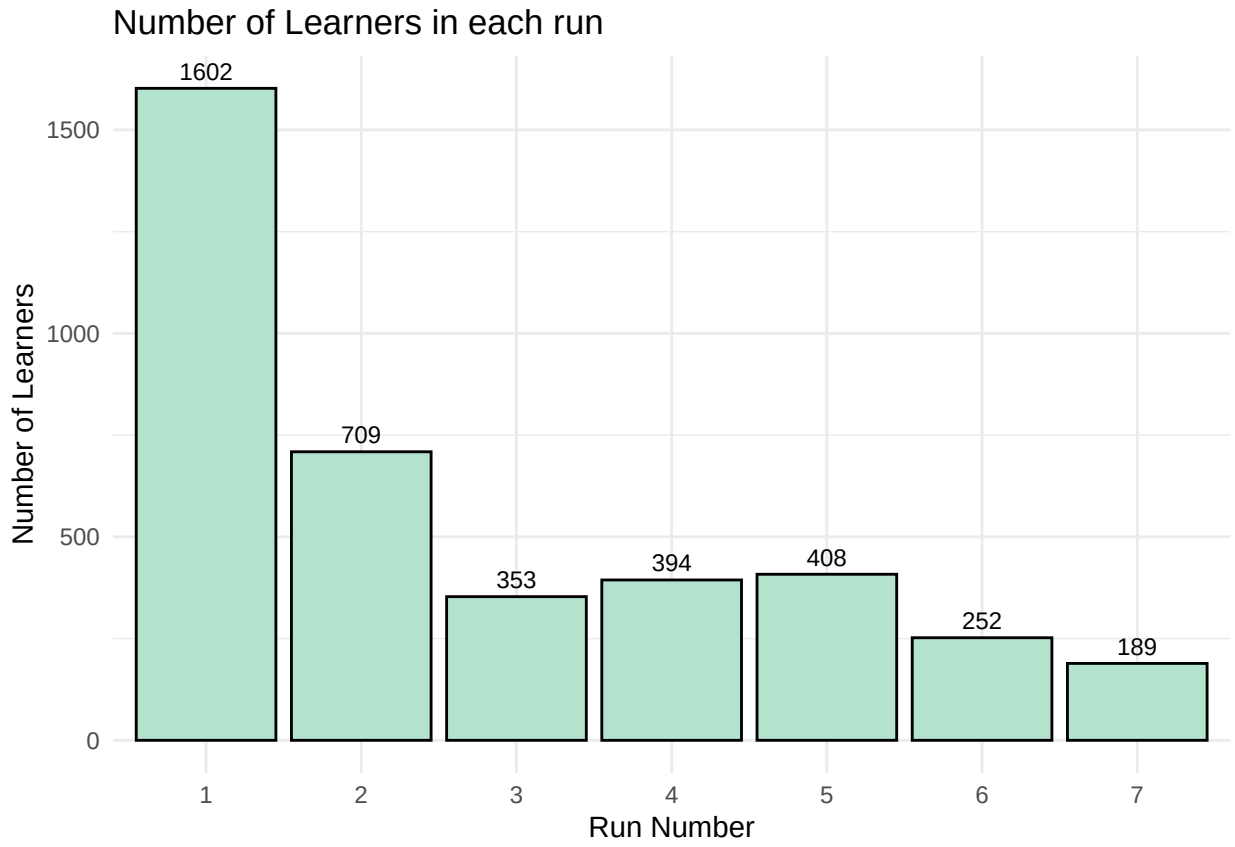
As I stated above, the data is not clean yet. Hence, in this step I cleaned all the null values and missing data and also filtered out the irrelevant columns. All the data cleaning is performed in the preprocessing script. After cleaning, the data kind of looks like this: (Only initial rows displayed.)

```
##                             learner_id gender country age_range
## 1 4dc22fed-63d4-4bf6-b162-bdf482e1ec38   male      PE     46-55
## 2 7a44b170-73f8-4863-8687-4f97934c8b0b   male      IT     36-45
## 3 3fc06ecd-3ef8-4f45-90e5-250c23c0ff71   male      GB     56-65
## 4 51c61184-822d-437a-b56e-a52a68da6e6e   male      GB     46-55
## 5 5b9b6f6c-db98-4424-ada8-977e654c94df   male      LS     26-35
##   highest_education_level employment_status                employment_area
## 1       university_degree working_part_time          teaching_and_education
```

```
## 2               secondary working_full_time accountancy_banking_and_finance
## 3       university_degree  looking_for_work         transport_and_logistics
## 4                tertiary     self_employed     it_and_information_services
## 5               secondary  looking_for_work                         Unknown
##   run_number
## 1          1
## 2          1
## 3          1
## 4          1
## 5          1
```

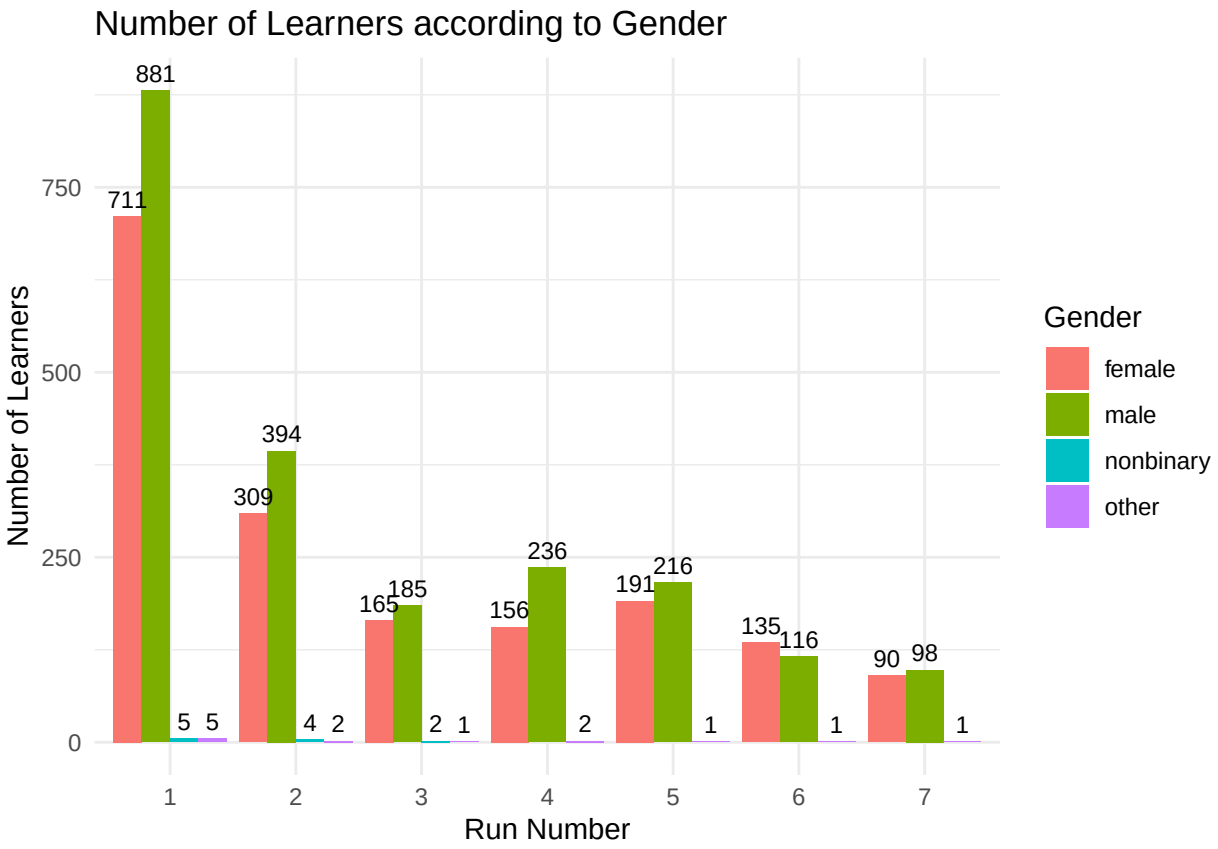
4. Modeling

This is one of the important phase of our cycle, where data analysis is done and results are drawn in the form of visualizations. According to my research question, to do demographic analysis I will make use of the demographic parameters like gender, age, etc. First, lets see the trend on popularity of the course for every run.



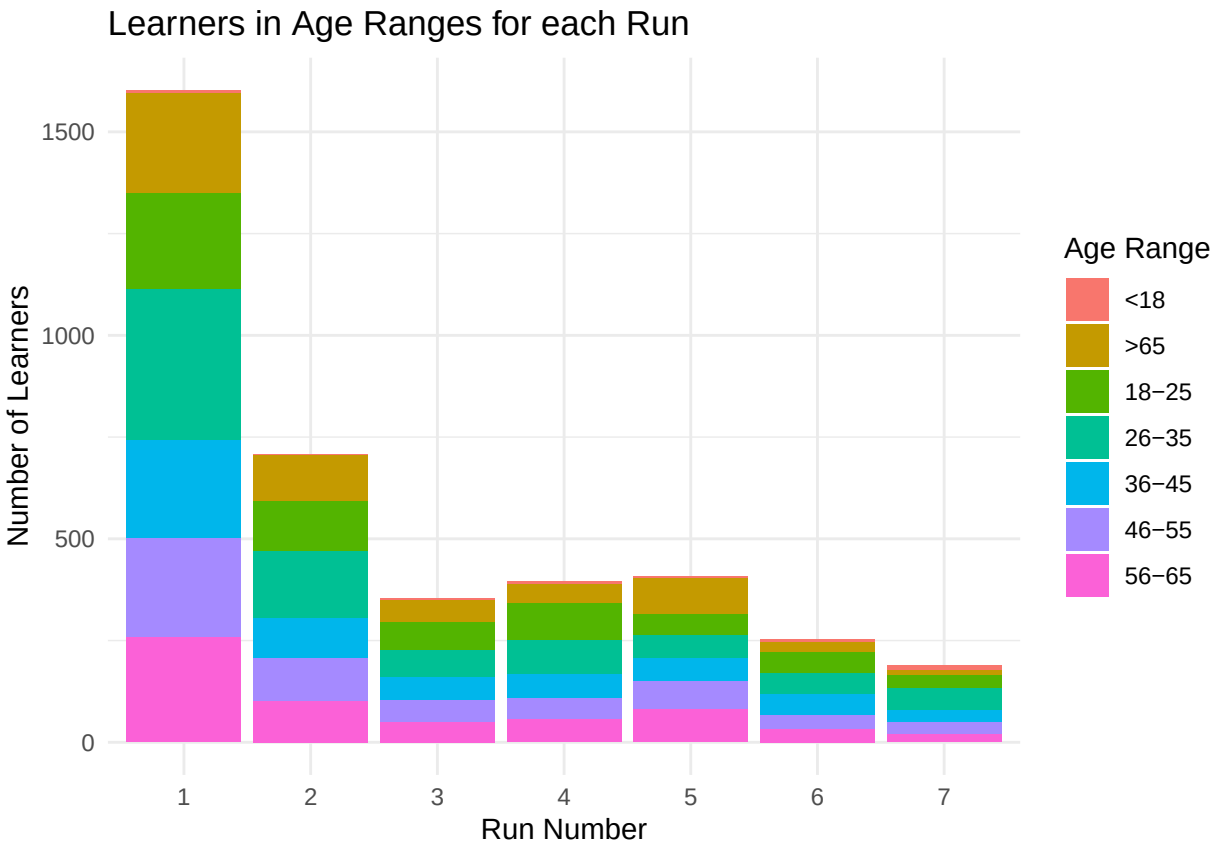
Insight - Above graph shows how the popularity of the course decreased in each run. The number of learners in the latest run is far more less than the first run. The providers should focus on the why there is less participation for the course and what can be done to improve it.

Next, I will move onto the demographic analysis with first analyzing the learners background based on gender. Below graph displays the number of learners in each run according to their respective gender.



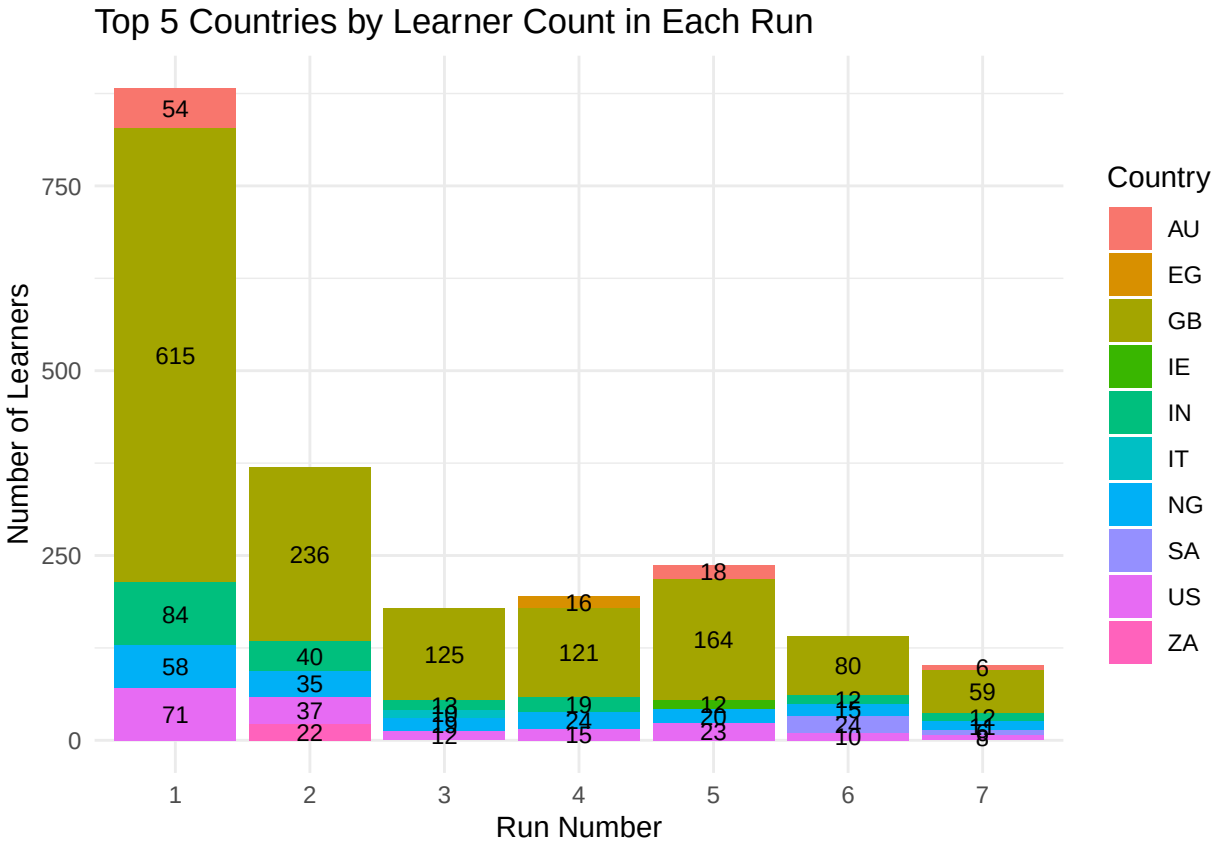
Insights - Here we can see that the difference between number of male and female learners in each run kept decreasing along the trend which is a good thing. The gender ratio is quite good and must be maintained for future runs.

Next, I analyzed the age ranges of learners and find out from what age range there is more engagement in the course for each run.



Insights - As we can observe the trend, the number of learners in all age groups have decreased over the runs. But the age groups from 18-25, 26-35, 36-45, 46-55 have shown almost equal number of learners in the later runs. The engagement of younger age is maintained but the same of older age is decreased. The providers might want to improve some factors to regain the involvement of older age range of learners.

Lastly, I visualized a trend of learners from where they belong. As the number of countries is more, I have considered the top 5 countries in each run.



Insights - As the graph shows, the engagement of learners from Great Britain is the most in each run. The engagement of international learners decreased gradually over all runs. The provider might consider to reach more international audience to increase their involvement.

5. Evaluation :

In this phase, I evaluate whether our analysis fulfilled the objectives and answered the question I raised. Also I will evaluate whether the success criteria is met, suggest next steps and improvements.

According to the EDA, I have analyzed and visualized the demographic variables relevant to our course. Also I have written out insights for every outcome which focuses on the changes in the trend over all runs of the course.

Recalling the business objectives, I have done the demographic profiling of the learners and introduced what is the background of the audience of the course to the provider. This answers our research question at a greater extent.

Also recalling the success criteria, I have done analysis which fulfills the business objective and answered the question raised. I have described the data quality as most of the data is irrelevant and inconsistent and have chosen only one that is useful for analysis.

After going back to the first analysis i.e. popularity of the course, in the next cycle, I'll be focusing on why learners are leaving the course in the middle and what can be done to increase the engagement of the learners.

CYCLE 2

1. Business Understanding :

The first phase of the CRISP - DM cycle is to describe the business understanding. In this, I will define the objectives and the success criteria of this investigation. My task in this cycle will be to clear the objectives by trying to answer the business question and meeting the success criteria by implementing required techniques.

1.1 Business Objectives :

Following to the last cycle, the business objective in this cycle will be : * To find out reasons affecting the engagement of learners in the course. * To find out parameters impacting the engagement of the learners. * To figure out solutions to the make relevant changes that will increase the engagement.

1.2 Success Criteria :

- The data used should be accurate and from the right source to produce relevant results.
- The implementation and insights should be able to complete the business objectives.
- The analysis should derive outputs that will help provide solutions to the problems and questions raised.
- The documentation must be clean and easily understandable to the providers in order to take actions needed based on our suggestions.

1.3 Data Mining Objectives :

- The data must be extracted from original dataset and in the right necessary formats.
- The data should be cleaned and preprocessed for analysis.
- Analysis done on the data should provide right solutions to our problems with the help of right choice of visualizations.

1.4 Research Question :

According to the problem raised in the last cycle, this cycle will try to answer the following question : “Why learners are leaving the course in the middle? What factors are affecting them to take this decision?”

2. Data Understanding :

2.1 Data Collection :

The data is collected from the dataset provided by FutureLearn of all runs of the course. There were separate set of data files for each run and in different formats. There were .csv files which had all the stats and data whereas, there were pdf files of each run which consists of how webpage of course of each run looks like. Here I am focused on the .csv files which will be relevant for my objectives. For our objective, I have chosen the “leaving-survey-responses.csv” files and enrollment data used in last cycle.

2.2 Data Quality :

As discussed earlier, I have chosen the enrollment data and leaving survey responses data for our objective. When I checked the files the quality of data was inconsistent. I have described the data quality of enrollment data in the first cycle so below is the description and the quality of each column from the leaving survey response data file. This is an overall review for leaving survey response files of all runs:

- id - Unique id for every survey. Irrelevant to our research.
- learner_id - Primary Id of learners who have unenrolled from the course.
- left_at - Time stamp when learners have unenrolled from the course. Data should be in Date format.
- leaving_reason - Reasons for leaving the course. Categorical and clean data column
- last_completed_step_at - Time stamp when learner completed the last step before leaving. Has missing values and inconsistent.
- last_completed_step - Last completed step by learner. Has missing values and inconsistent.
- last_completed_week_number - Last completed week of course. Has missing values and inconsistent.
- last_completed_step_number - Last completed step number by learner. Has missing values and inconsistent.

3. Data Preparation

This phase focuses on the data preprocessing techniques starting from the selection of right filtered data for our research. I will perform steps like collecting, integrating, cleaning and preprocessing the data in this phase. All the major preprocessing of the data is performed in the preprocessing script in the 'munge' folder of the project directory. This phase provides the cleaned and filtered data to create different visualizations for our EDA which will be discussed in next phase.

3.1 Data Selection :

First as discussed in data collection, I imported the data into dataframes and integrated them together by adding run_number column to each row which stated from which run the data of learner is. After merging the data, I calculated the number of learners who left in each run to kickstart a trend.

```
## # A tibble: 4 x 2
##   run_number learners_left
##        <int>         <int>
## 1          4            67
## 2          5           173
## 3          6            83
## 4          7            80
```

3.2 Data Cleaning :

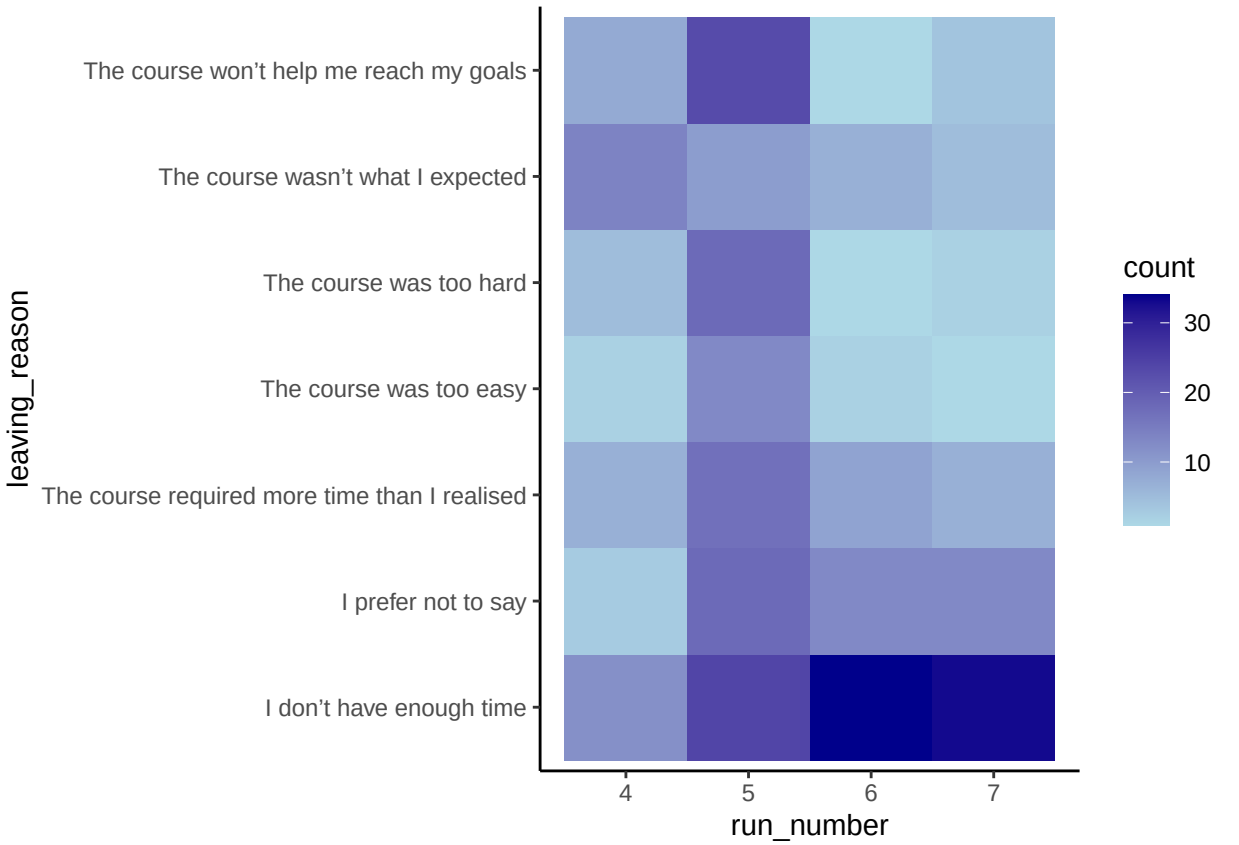
In this step I filtered the leaving data which had a value in leaving reason as "Other" which is irrelevant for my research question. I won't be cleaning any missing rows of columns which are not relevant to the research question. Our main focus is on the leaving_reason.

```
##     id                              learner_id                 left_at
## 1  212 cbeff0e9-104a-4b4e-ab8d-0e9d4a73f02e 2017-11-24 16:05:42 UTC
## 2  747 7f52aeb2-68b1-4d89-b167-d6b5a104138b 2017-11-25 10:52:20 UTC
## 3  785 a153788d-39f2-40f1-a360-ab944ee612bb 2017-11-25 12:33:05 UTC
## 4 2831 5cbc8827-23de-415c-aa41-b3f2c9f28c0f 2017-11-27 17:51:03 UTC
```

```
## 5 2948 efbd0ec1-d3f2-490f-9a21-b423ff9e4e93 2017-11-27 20:01:27 UTC
## 6 4681 74082dd9-2f29-4d39-a050-f76849f92747 2017-11-29 14:33:12 UTC
##                                 leaving_reason  last_completed_step_at
## 1                     I don't have enough time
## 2                          I prefer not to say 2017-11-25 10:45:28 UTC
## 3                     I don't have enough time
## 4            The course wasn't what I expected 2017-11-20 18:01:15 UTC
## 5                     I don't have enough time
## 6 The course required more time than I realised
##   last_completed_step last_completed_week_number last_completed_step_number
## 1                  NA                         NA                         NA
## 2                3.20                          3                         20
## 3                  NA                         NA                         NA
## 4                1.16                          1                         16
## 5                  NA                         NA                         NA
## 6                  NA                         NA                         NA
##   run_number
## 1          4
## 2          4
## 3          4
## 4          4
## 5          4
## 6          4
```

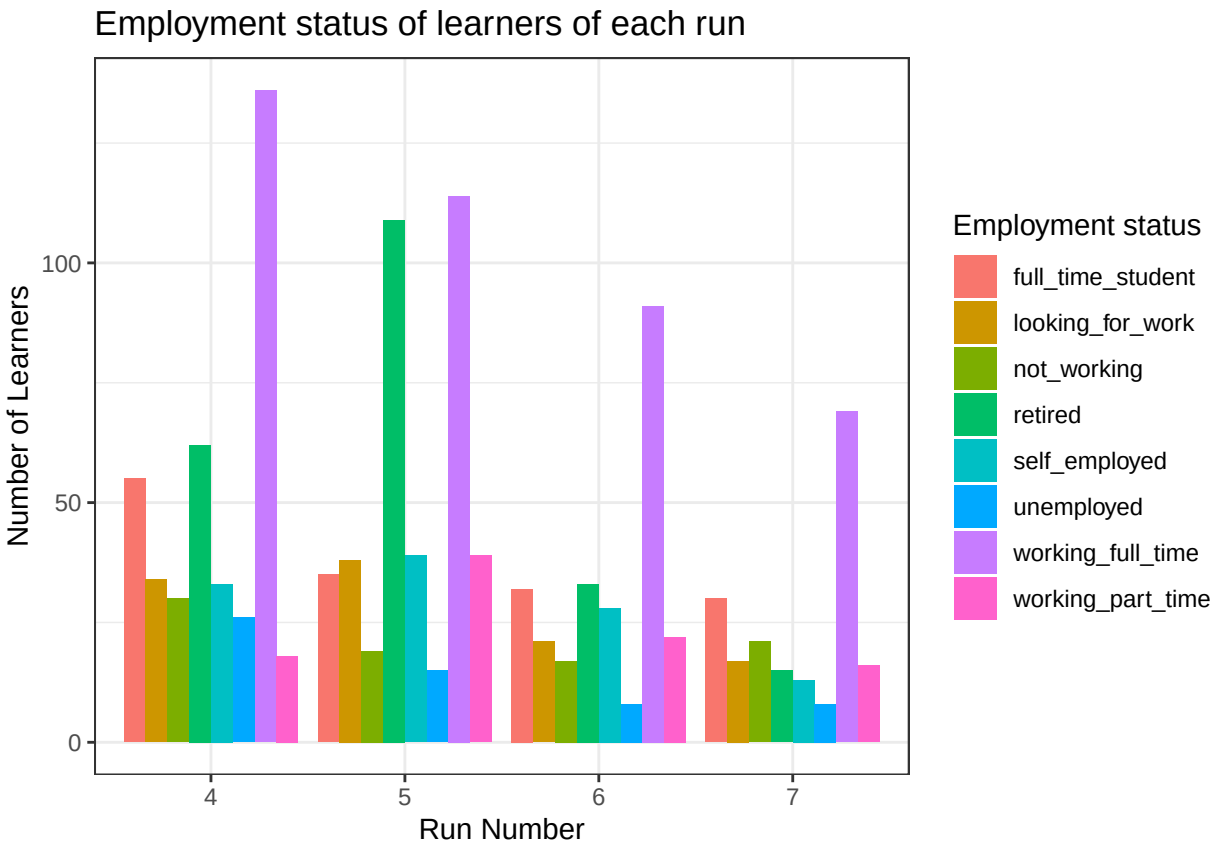
4. Modeling

This is one of the important phase of our cycle, where data analysis is done and results are drawn in the form of visualizations. According to my current research question, I will be analyzing the reasons for which learners have left the course.



Insights - From the heatmap we can see that in the latest runs, most of the learners left the course because they don't have enough time. There was a mixed bag of reasons in 5th run. The provider must tackle this problem by either providing more time flexibility to the learners or altering the course content or one of my best suggestions would be to let the learners create their own paths by choosing the only topics they want to learn or focus more on tests and quizzes.

After inspecting the data, I think that employment status of the learners might have impact on the reasons.



Insights - As we can see above, the most of the learners in all runs are working full time. And according to latest runs, the ratio of learners working full time is far more than other statuses. The learners working full time might not be having enough time to engage in the course. Providers might want to make the course more flexible so that all kinds of learners can remain engaged.

5. Evaluation :

In this phase, I evaluate whether our analysis fulfilled the objectives and answered the question I raised. Also I will evaluate whether the success criteria is met, suggest next steps and improvements if any.

According to the EDA, I have analyzed and visualized the leaving surveys of our course. Also I have written out insights for every outcome which focuses on the changes in the trend and the impacting reasons and factors over all runs of the course.

Recalling the business objectives, I have found out the reasons for lesser engagement of learners and also the factors impacting their decision. This answers our research question at a greater extent.

Also recalling the success criteria, I have done analysis which fulfills the business objective and answered the question raised. I have described the data quality as most of the data is irrelevant and inconsistent and have chosen only one that is useful for analysis.

In this way, our analysis has satisfied all the business objectives in both the cycles.

Deployment :

This is the last phase of our both the cycles, where I will present my results and insights to the providers. I have created this report and a presentation to showcase my research and deliver the significant insights

drawn from my analysis. I intend to include my visualizations in the presentation to best explain my analysis to the providers. Also, I have stated the question raised in the research and also the suggestions for the same. This will give them an ideal intake of my research in the easiest way.