Cyber Security: Safety At Home, Online, and in Life

Varsha Venkataraman - 240150459

2024-11-10

Introduction

This report investigates and provides insights on the dataset collected for the Cyber Security: (Safety At Home, Online, and in Life) course developed by Newcastle University and offered by the Future Learn Platfrom which is a massive open online course (MOOC). The course is 3 weeks long and was free to access.

Business Understanding:

The university course team are likely to find some value from this report, with the main beneficiaries being the online course providers. The framework that is used to provide the results and conclusions in the report is the Cross Industry Standard for Data Mining (CRISP-DM), performing two cycles of the model to produce the results which follow.

The investigation can provide them in framing the upcoming course structure or modify the existing ones, and rectify the existing issues they are facing with learners.

Success Criteria:

The main objective of the report is to identify the areas where the university can improve in getting the right choice of learners and improve the course modules so that the course will also be popular and effective.

Round 01 of the CRISP-DM Cycle

With the primary objective in the mind the first question which this report aims to answer is:

Research Question: What is the distribution of employment areas among learners enrolled in the Cyber Security course, and are there any missing data which is in high proportion that affect the insights?

1.1 Data Understanding:

For phase 1 of CRISP-DM,the data set that has the details of learners such as the learner id, employment area, are the primary columns required. The cyber security enrollments file for each period are considered primarily for the first research question.

1.2 Exploring the data:

After gathering the dataset the next step is to investigate for any data issues. This step is to avoid any potential issues that might arise when we start with the analysis part, Hence its necessary to clean and pre process the dataset as per our requirement.

When I explored the cyber security enrollment files for the cyber security for different time frames. The dataset contained learners Id and their specific 'employment_area', 'highest_education_level' and few more details. But, there were lots of 'Unknown' values in many rows and columns. The 'Unknowns' values created the data to be more weaker to come up with a conclusion as these values cannot be assumed or filled with methods to fill the missing entries.

Few columns had 'Unknown' as a value , but column like 'purchased_statement_at' have many missing data in it and the rows were also left blank. Categorical fields like these are difficult to be filled with data, and we need to consider them to be 'NA' or 'Unknown'.

Also, there were repeated learners id , initialy I considered it to be a primary key but when I saw it to be repeated, a doubt of duplicates araised. But, further exploring the repeated values along with its enrolled date, a clear understanding was made that the learners rejoined the course at various intervals.

1.3 Data Preparation:

The data preparations included five steps in total. The steps includes data collection, data cleaning , data integration, data transformation and selecting the data in preparation for the modelling phase to come next.

1.3.1 Data Wrangling:

1.3.1.1 Data Munging:

The data gathered for this research was sourced from Future Learn. The Learner IDs allow for combining of information between different data sheets. The data is collected for the cyber security course for a specific period so there seven different time frames in which the data is collected.

The data has all the primary deatils of the learners who joined the course in different time frames, and their specific enrollment time along with their employment area is also gathered.

The `munge` folder contains files related to data pre-processing steps. This folder includes:

```
-> munge
   |_ 01-dataCleaning.R
   |_ 02-dataIntegration.R
   |_ 03-dataTransformation.R
```

1.3.1.2 Data Cleaning:

Data wrangling is a more specific part of data preparation, focusing on cleaning, restructuring, and transforming raw data into a more useful format whic includes the below steps.

1.3.1.2.1 Data Filtering:

In data cleaning, I filtered out the required columns alone and removed the unwanted columns from the dataset. In the enrollment file , I only considered 'learner_id', 'highest_education_level', 'employment_area' and eliminated rest of the columns.

1.3.1.2.2 Data Reshaping:

The 'enrolled_at' column had the format of date and time, I filtered out only date from it and reshaped the row value.

1.3.1.2.3 Handling Missing Values:

I also replaced the empty rows with 'NA' for uniformity and to fill the blank space. Here we are not able to follow any of the methods such as to replace with mean , mode or median for the missing datas as all are unique values and its hard to follow any of the existing methods to replace hence went with the way of filling it with the unknown value.

Why 'NA' instead of 'Unknown' :

NA is a standardized way to represent missing data in R (and other statistical tools). Most functions that handle missing data recognize NA, making data processing more seamless. Using "Unknown" as a placeholder may lead to misinterpretation by functions that expect missing values as NA. Most statistical models, transformations, and data wrangling techniques treat NA values appropriately, either by ignoring or imputing them as necessary. Having "Unknown" in place of NA might bias analyses, as it's recognized as text rather than as a missing value.

1.3.1.3 Data Integration:

In data Integration, In figure 1, there are different files which are pre processed and data is cleaned, Now I combined the related files altogether into a single file for each group. So, the figure 2 represents the combination of all related files.

```
-> cache
   |-> preData
         |_ cleaned_cyber_security_1.csv
         |_ cleaned_cyber_security_2.csv     -> cache
         |_ cleaned_cyber_security_3.csv         |-> finalData
         |_ cleaned_cyber_security_4.csv               |_combined_cleaneded_cyber_security.csv
         |_ cleaned_cyber_security_5.csv
         |_ cleaned_cyber_security_6.csv
         |_ cleaned_cyber_security_7.csv
```

1.3.1.4 Data Transformation:

In this part, I re-coded the employment_area row values with standardized categorical labels. This transformation is usually done when you want to standardize or simplify categorical values, making them easier to analyze or interpret.

Old Value	New Value	Old Value	New Value
accountancy_banking_and_finance	Finance	law	Law
armed_forces_and_emergency_services	Defense	marketing_advertising_and_pr	Marketing
business_consulting_and_management	Consulting	media_and_publishing	Media
charities_and_voluntary_work	Charity	property_and_construction	Construction
creative_arts_and_culture	Arts	public_sector	Public
energy_and_utilities	Energy	recruitment_and_pr	Recruitment
engineering_and_manufacturing	Engineering	retail_and_sales	Retail
environment_and_agriculture	Environment	science_and_pharmaceuticals	Science
health_and_social_care	Healthcare	teaching_and_education	Education
hospitality_tourism_and_sport	Hospitality	transport_and_logistics	Logistics
it_and_information_services	IT		

1.3.1.5 Feature Engineering:

For the first research question, its important to have learners id,and to find distribution of employment area of the learners we need to consider the employment area feature. So rest fields are optional for this research question hence removed them for further analysis.

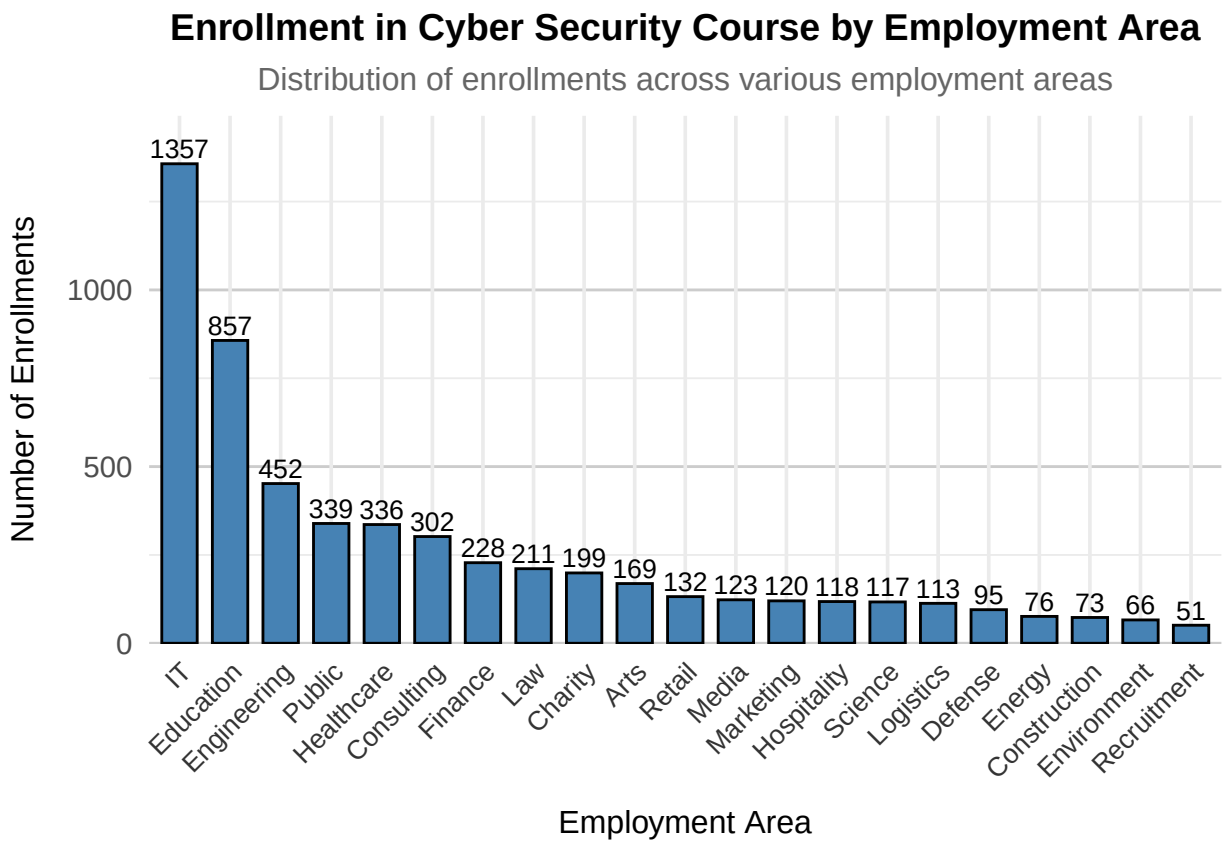
1.4 Modelling and Visualizations:

The below first bar plot 'Enrollment in Cyber Security course by Employment Area' represents the 'Employment Area' of the learners who enrolled in the cyber security course in the university. The y axis has the different type of 'Employment Area' of the learners. The x axis has the number of enrollments in the cyber security course which has the range from 0 to 1500.

From the visualization we can conclude that the learners from IT industry has the maximum enrollment and the learners from the recruitment industry has the least enrollment. It is very obvious that the IT field is related to the course, but its surprising that the Finance area is standing seventh in the column.

The cyber security has important role in the finance and banking area as many fraudulent acts occurs in this area.

The second bar plot titled 'Number of rejoiners by Employment Area' depicts the maximum number of rejoiners from different employment area. Its very evident that its from 'IT' field there are maximum number of rejoiners. Also, the number of rejoiners is not propotional to the number of enrollments in the course base on employment area.


Enrollment in Cyber Security Course by Employment Area
Distribution of enrollments across various employment areas
Number of Enrollments
0
500
1000
1357
857
452
339
336
302
228
211
199
169
132
123
120
118
117
113
95
76
73
66
51
IT
Education
Engineering
Public
Healthcare
Consulting
Finance
Law
Charity
Arts
Retail
Media
Marketing
Hospitality
Science
Logistics
Defense
Energy
Construction
Environment
Recruitment
Employment Area

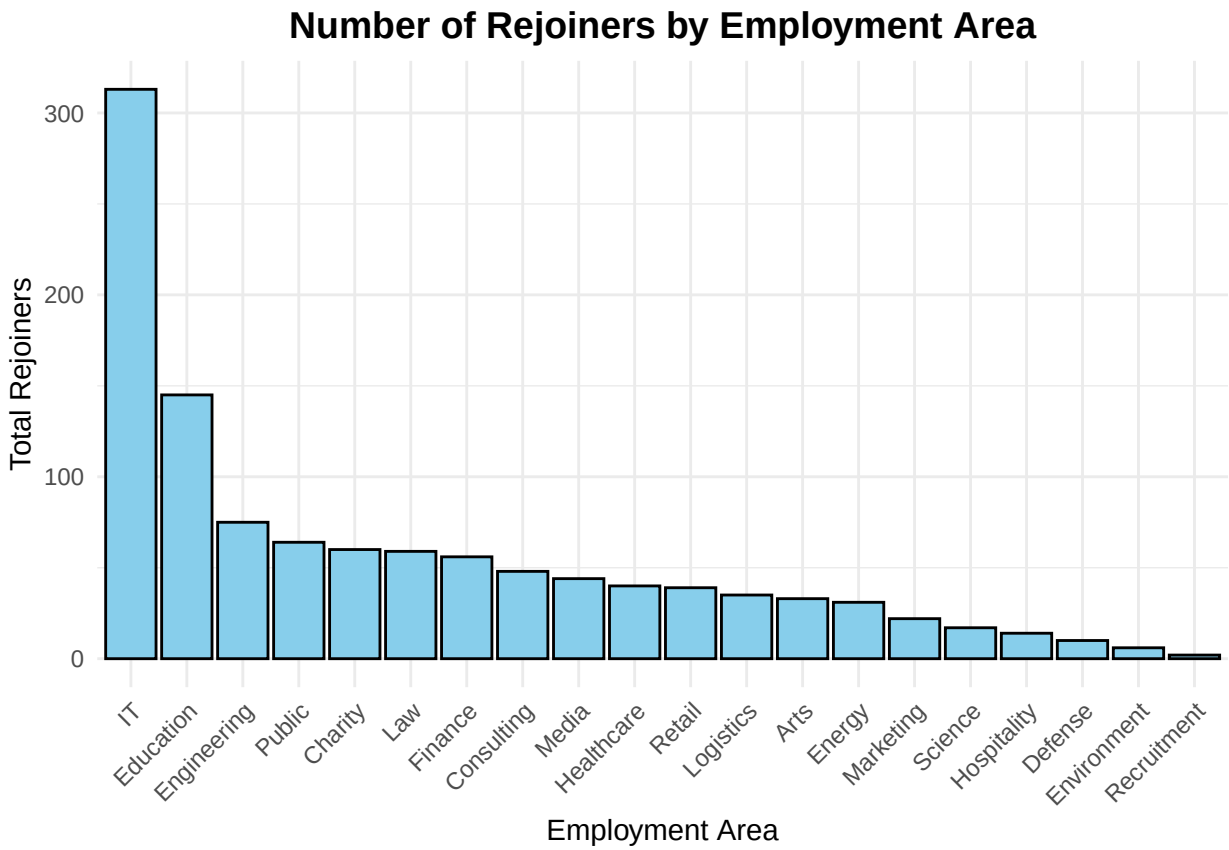

Number of Rejoiners by Employment Area
Total Rejoiners
0
100
200
300
IT
Education
Engineering
Public
Charity
Law
Finance
Consulting
Media
Healthcare
Retail
Logistics
Arts
Energy
Marketing
Science
Hospitality
Defense
Environment
Recruitment
Employment Area

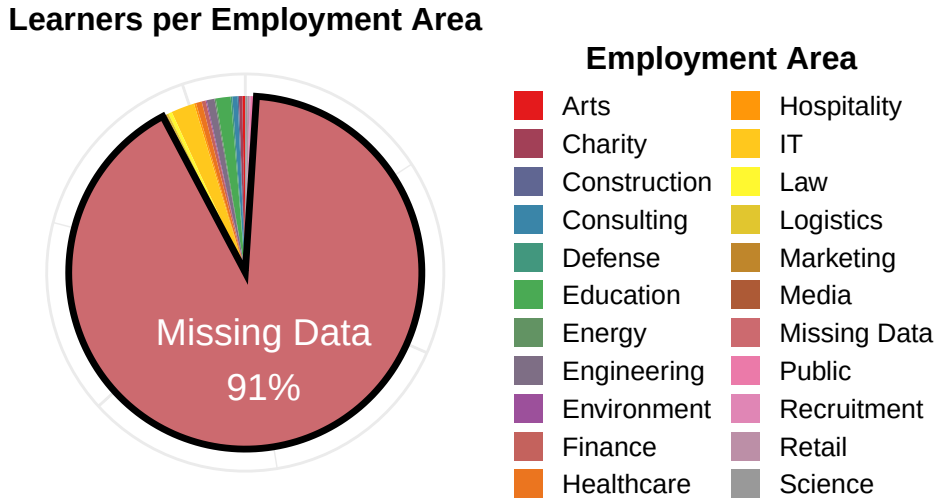
Table 2: Learner Metrics

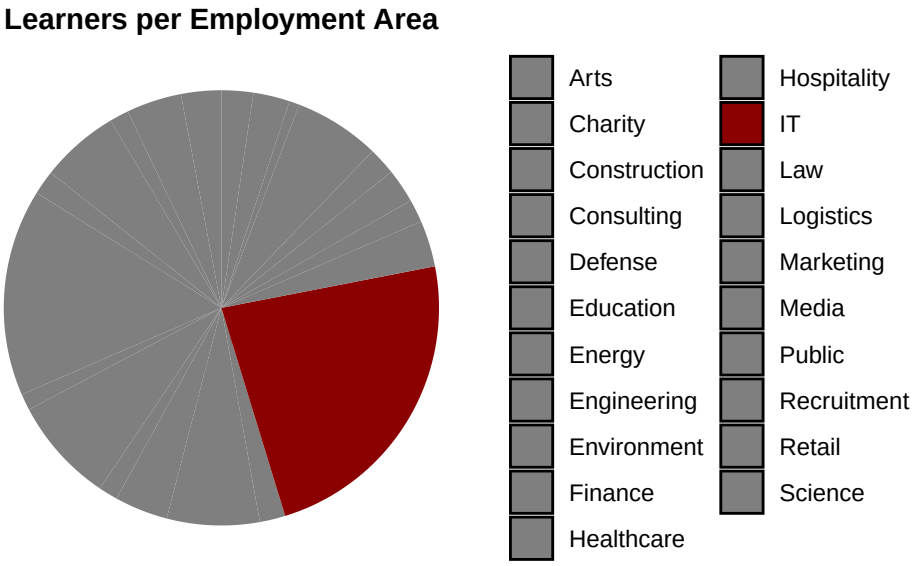
Metric	Value	Metric	Value
Total Learners	63248	Unique Learners	35225

With respect to the above Table 1, the total number of learners who are enrolled in the cyber security course is 63,248 of them. Looking at the pie chart below which has the percentage value for the learners per employment area. We can conclude that we have the employment area information only for 9% of the learners only.

The remaining 91% (roughly 57,556 learners) do not have employment area information. Because of which we cannot come up with 100% accuracy which area of employers take up this cyber security course.

There are around 63,248 learners with respect to the Table 1, who enrolled altogether in the cyber security course. But the unique number of learners are only 35,225 according to Table 1. Which means the rest 28,023 are either duplicate or re-joiners.





1.3.1.2 Evaluation :

The goal of first research question is not completely achieved as there are 91% of missing data so we cannot come to a conclusion with cent percent accuracy that the IT industry has the most number of learners for

the cyber security course. With the dataset we have excluding the missing data, we can make a conclusion that the IT industry has the maximum distribution of learners.

We need to ensure to have the enrollment details completely collected before enrolling and starting with the course, as it is always necessary to have the basic enrollment details of the learners.

Round 02 of the CRISP-DM Cycle

With the secondary objective of the report, I am investigating on number of re-joiners, I am trying to understand the reason for them to quit and rejoin. Also, to gain some insights so that we can avoid re-joiners and modify the course modules accordingly to make it efficient to complete in the first schedule, if the reasons are stated because of the course structure.

Research Question: What is the primary reason for re-enrollment and which employment area has the majority re-joiners?

1.1 Data Understanding:

For phase 2 of CRISP-DM, to make an analysis for the number of re-joiners I first need to know the count of re-joiners and their duplicate learners id. Also, to know the primary reason to leave the course its required to gather the unique reasons stated by the learners to quit. So for this I am considering the leaving_survey_response data and also the course enrollment data.

1.2 Exploring the data:

Firstly, the course enrollment data is analysed in the first cycle and number of unique and repeated learners count is gathered. Looking into the leaving_survey_response data we can gather the unique reasons for the learners to leave the course.

1.3 Data Preparation:

1.3.1 Data Wrangling:

1.3.1.2 Data Cleaning and Filtering:

In data cleaning, I filtered out the required columns alone and removed the unwanted columns from the dataset. In the survey response file , I only considered 'learner_id', 'left_at', 'leaving_reason' and eliminated rest of the columns.

1.3.1.2.2 Data Reshaping:

For the duplicate learners id , the 'enrolled_at' column had multiple dates in it, so i had to split each dates into separate columns for better analysis. I created a new column 'unique_enrolled_at' and splited multiple dates into separate ones. Now, if the learner had rejoined for more than three times to the course it would be three different columns so that we can understand just by looking into the columns and it is also to analyse the time frame.

1.3.1.2.3 Handling Missing Values:

Moreover the first three leaving-survey-responses files did not even have any datas the file size was less in bytes. So, I did not consider these files and started with the files that had data.

I also replaced the empty rows with 'NA' for uniformity and to fill the blank space. Here we are not able to follow any of the methods such as to replace with mean , mode or median for the missing datas as all are unique values and its hard to follow any of the existing methods to replace hence went with the way of filling it with the unknown value.

Why 'NA' instead of 'Unknown' :

NA is a standardized way to represent missing data in R (and other statistical tools). Most functions that handle missing data recognize NA, making data processing more seamless. Using "Unknown" as a placeholder may lead to misinterpretation by functions that expect missing values as NA. Most statistical models, transformations, and data wrangling techniques treat NA values appropriately, either by ignoring or

imputing them as necessary. Having "Unknown" in place of NA might bias analyses, as it's recognized as text rather than as a missing value.

1.3.1.3 Data Integration:

In data Integration, In figure 1, there are different files which are pre processed and data is cleaned, Now I combined the related files altogether into a single file for each group. So, the figure 2 represents the combination of all related files.

```
-> cache
   |-> preData
         |_cyber_security_leaving-survey-responses_4.csv
         |_cyber_security_leaving-survey-responses_5.csv
         |_cyber_security_leaving-survey-responses_6.csv
         |_cyber_security_leaving-survey-responses_7.csv
-> cache
    |-> finalData
        |_combined_cyber_security_leaving-survey-responses.csv
```

1.3.1.4 Data Transformation:

The column 'enrolled_at_' is in the format 2018-07-06 10:55:39 UTC, but i wanted only the date to understand the time range so i transformed it and filtered only the date '2018-07-06'. And, the 'enrolled_at_' had multiple entries like for an example '2018-07-06 10:55:39 UTC, 2018-08-06 10:55:39 UTC' , so i filtered only the date part even if it had multiple values and created another column which i mentioned in the Data Reshaping and stored only date in each rows.

1.3.1.5 Feature Engineering:

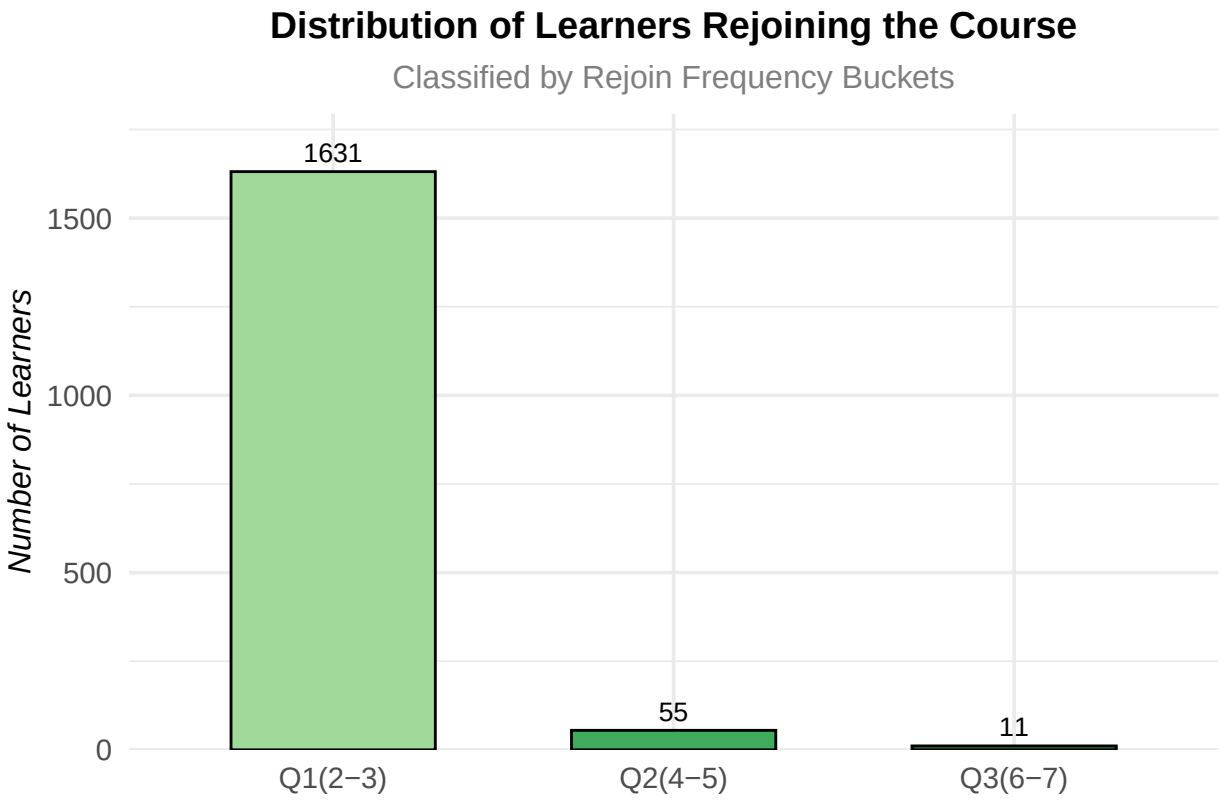
For this research question, its important to have learners id,and to find the reason for them to leave we need to have the 'leaving_reasons' field. So rest fields are optional for this research question hence removed them for further analysis.

1.4 Modelling and Visualizations:

To start with the visualization, the first bar plot title 'Distribution of learners rejoining the course' here I bucketed the count of re-joiners as groups, (Q1 has the count 2 and 3 re-joiners, Q2 has the count of count 4 and 5 re-joiner, Q3 has the count of 6 and 7 re-joiners). The x axis has the number of rejoiners count and the y axis has the number of learners. Looking at the plot we can come to a conclusion that there are maximum 2-3 re-joiners for the course.

Since there are many number of re-joiners for the cyber security course. Now, looking at the survey response collected from the learners for leaving the course I have tabled number of learners reasons to leave the course with unique reasons and its total count chosen by the learners.

Looking at the below tabular column I can see the maximum number of learners had opted saying 'They do not have enough time'. Also, 40 learners opted for 'The course required more time' since both are time related I added the values and the count is 143. There are learners who are not satisfied with the course and few found it hard and few found it very easy but there count are very less when compared to learners who choose time as main issue.



Unique Leaving Reason	Leaving Reason Count
I don't have enough time	**103**
Other	97
I prefer not to say	47
The course required more time than I realised	40
The course wasn't what I expected	36
The course won't help me reach my goals	36
The course was too hard	26
The course was too easy	18

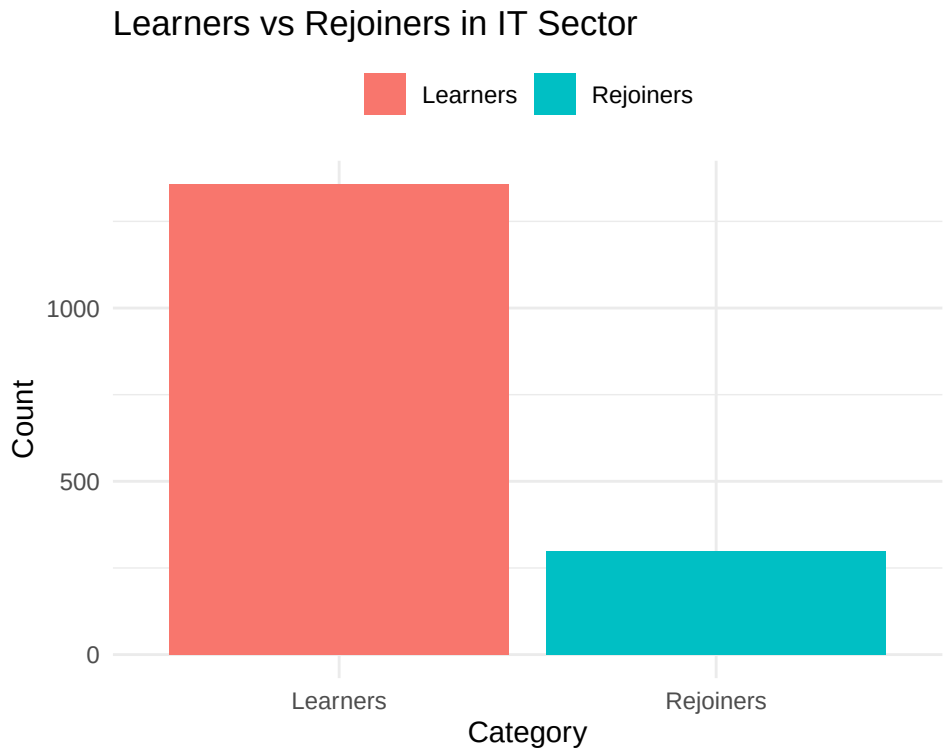
Cyber Security Level	Total Duration (hours)
cyber security 3	**50.05**
cyber security 4	**50.05**
cyber security 5	**50.05**
cyber security 6	**50.05**
cyber security 7	**50.05**

Since, the maximum reason was time constraints, I calculated the total hours taken for the course to be completed it is 50.05. Hence one main action by the course team need to take into consideration is to reduce the total time taken for the completion. Keeping in mind there are a lot of missing datas, we need to specifically see the reason to leave the course by the learners particularly from each employment area.

Looking at the below table, we can see the 'Arts' , 'Consulting' ,'Education', and 'Public' field related learners are feeling time related issues. Adding to it, we inffered that 'IT' field learners are more learners and re-joiners for the course so to see specific reasons choosen by the learners will also add value to the course team to take neccessary actions.

Leaving Reason	Employment Area	Count
I don't have enough time	Arts	33
I prefer not to say	IT	9
Other	IT	9
The course required more time than I realised	Arts	2
The course required more time than I realised	Consulting	2
The course required more time than I realised	Education	2
The course required more time than I realised	Public	2
The course was too easy	Education	2
The course was too easy	Marketing	2
The course was too hard	IT	13
The course wasn't what I expected	IT	11
The course won't help me reach my goals	Law	7

Looking at the above table we can see that, The learners from 'IT' employment area had choosen reasons like 'The course was too hard' , 'The course was not upto their expectation' , 'They prefer not say' and 'Other'. A suggestion to the course team would be if the learner is choosing 'other' to give an option to the learners to fill the reason so that it will help to know the exact reason. But with the current available data a decision that the course modules needed to be modified so that the learner from the 'IT' field will have complete satisfaction.



In conclusion it is stated that most of the learners feel that the course is very lengthy that they do not find enough time to complete. Which is why they are re-joining it or maybe the course is framed perfect its the learners are not able to frame time for learning.

If this total course time is minimized or divided as 25 hours for video materials and the rest modules as pdf or document to self study it might minimize the learners to leave the course.

Since, more learners are from 'IT' field we can collect the form as why the course is hard and why it did not

meet their expectations with which we can proceed to alter the modules.

1.3.1.2 Evaluation :

1.3.1.2 Deployment: