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

Varsha Venkataraman - 240150459

2024-11-09

Introduction

This report investigates and provides insights on the data set 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 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.

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:

```
-> 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
```

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.

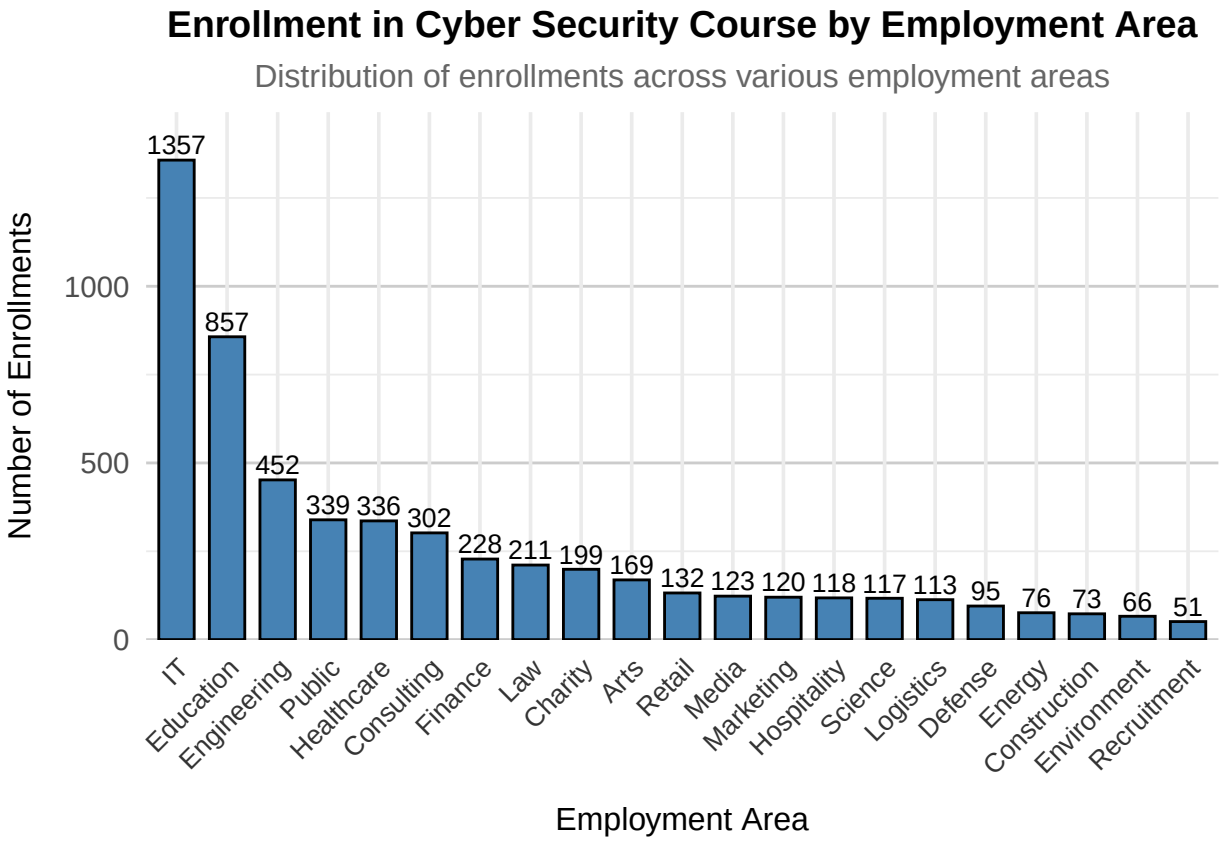
1.3.1.4 Data Transformation:

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:

1.4 Exploratory Data Analysis and Visualizations:



The above bar plot 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.

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.

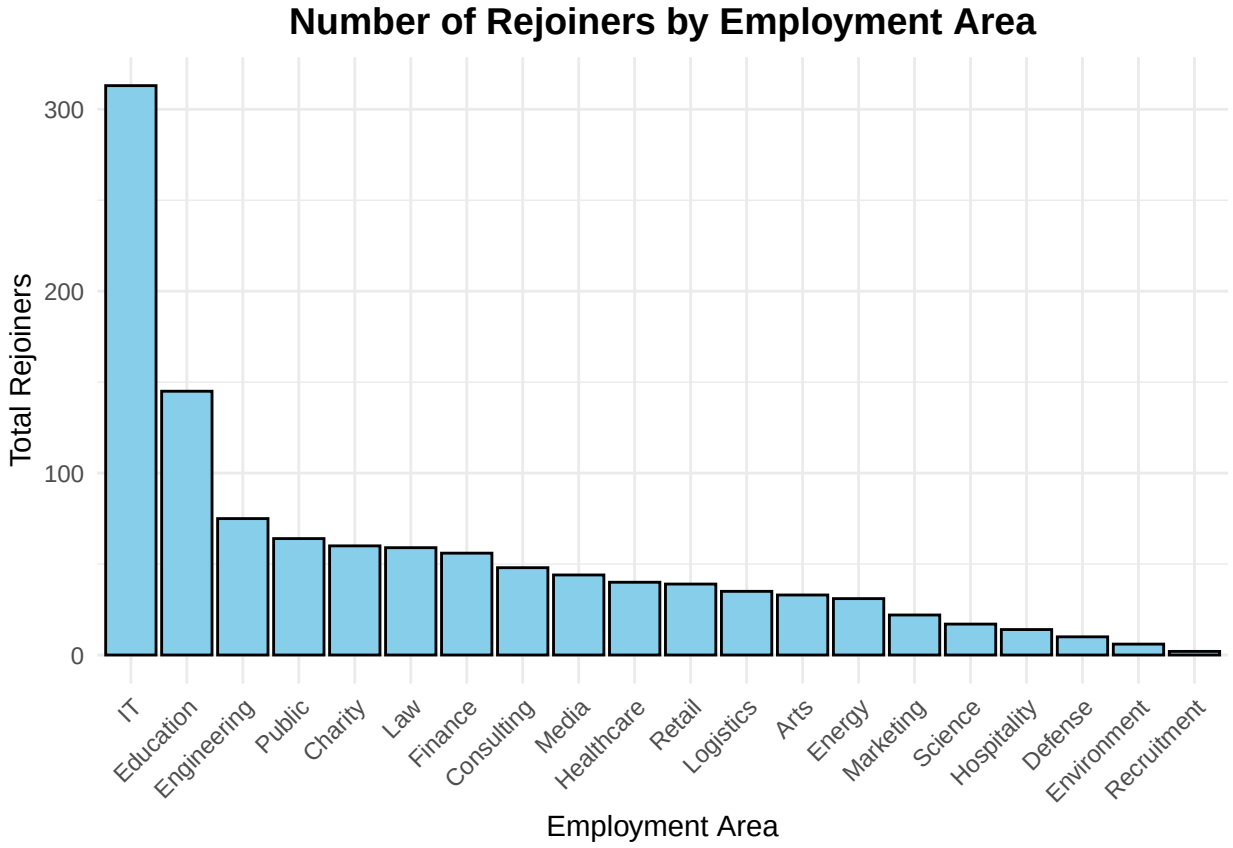
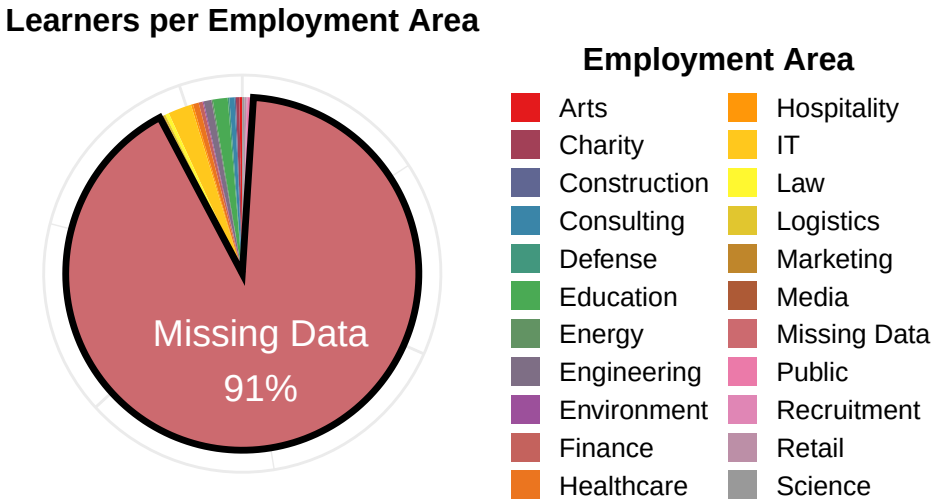


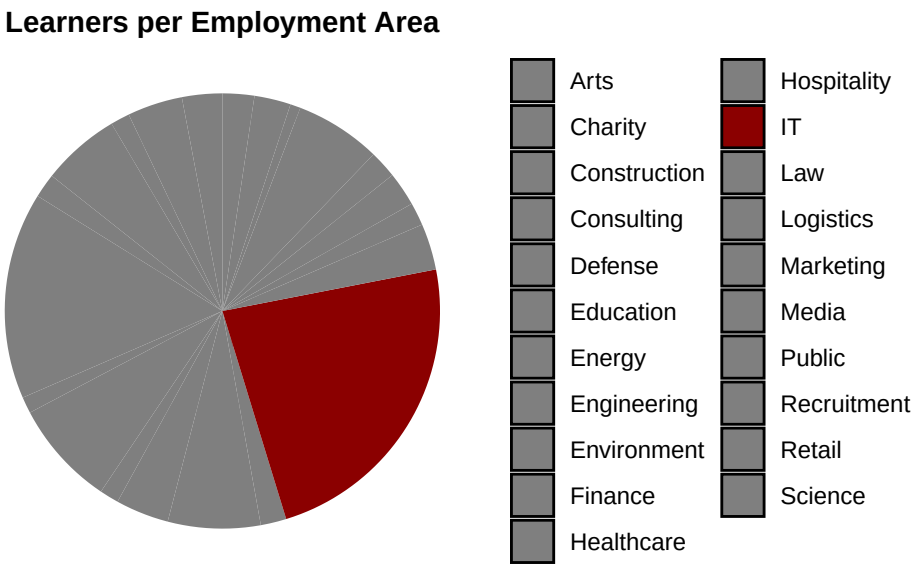
Table 2: Learner Metrics

Metric	Value	Metric	Value
Total Learners	63248	Unique Learners	35225

The remaining 91% (roughly 57,556 learners) do not have employment area information. Because of which we cannot come up with 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.





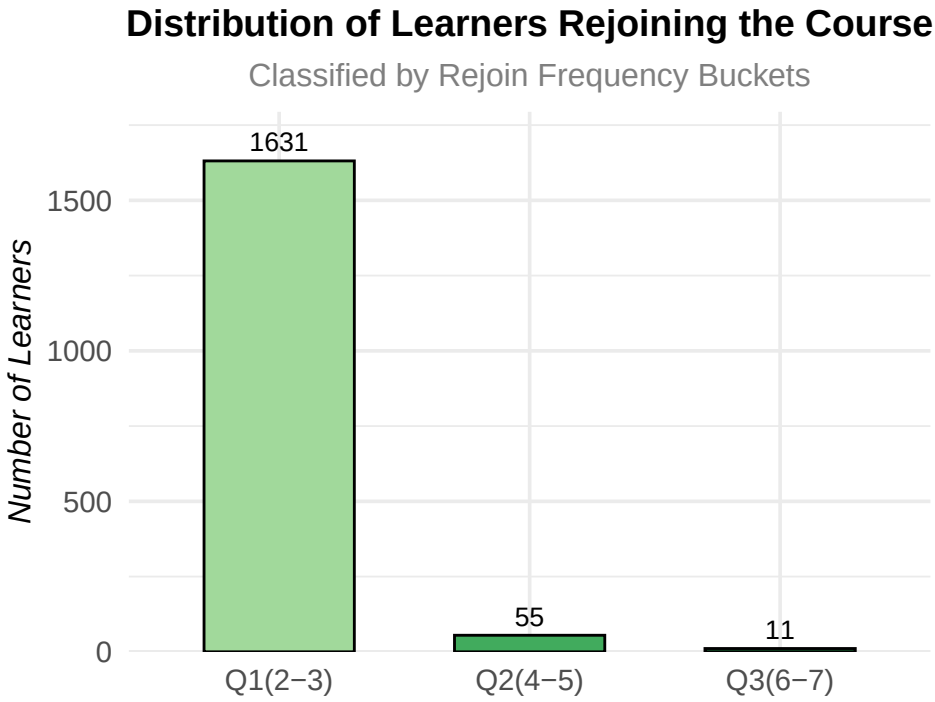
So to investigate on this I am trying to plot a visualization where I have grouped the re-joiners 2-4 as one bucket and 5-8 rejoining count learners as another bucket. And from the visualization below we can conclude that there are more number of re-joiners who are coming under the bucket count 2-4.

Round 02 of the CRISP-DM Cycle

Research Question: What is the primary reason for re-enrollment, related to the total course hours required?

1.1 Data Understanding:

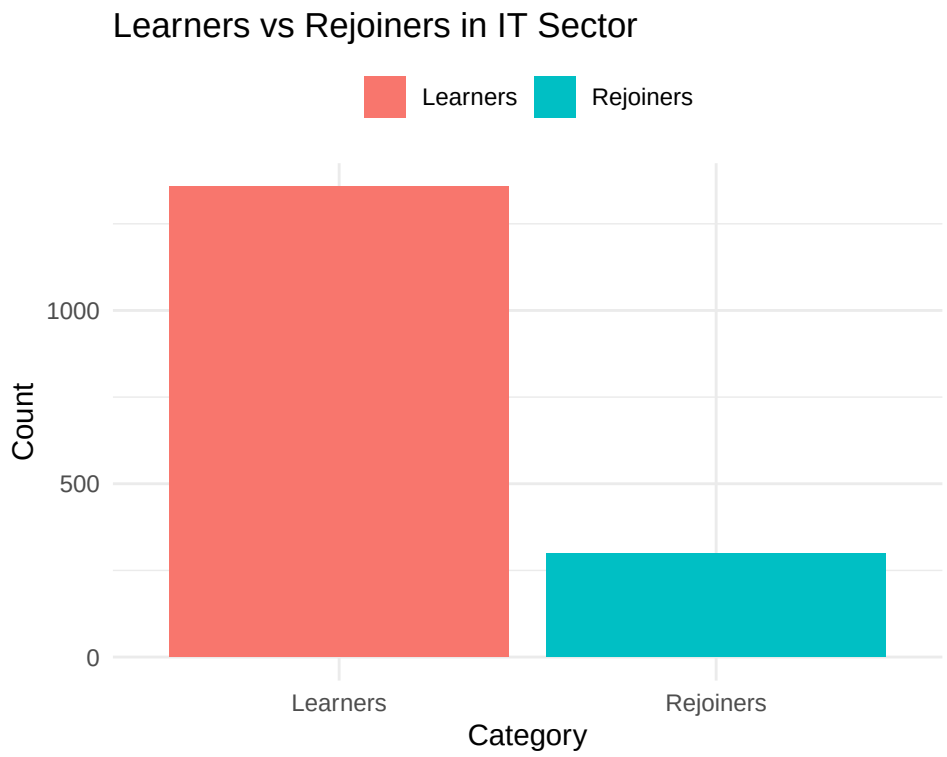
For phase 1 of CRISP-DM,the data set that has the details of learners such as the learner id, enrollment date, and some more information is required.The cyber security enrollments file for each period is to be considered.



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.

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**



From the above table 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.

Nearly 144 learners selected 'I prefer not to say' and 'Others' and 40 of them selected 'The course required more time than I realized' which altogether sums up 359 of them feel like time is the problem. Only 72 learners expectations did not match with the course and 26 of them found it hard.

Looking at the reasons for the learners to leave the course, The below tabular has the details of number of hours the course counts in total, so with respect to this we can conclude the total time taken by the course modules is 50.05 in hours.

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