Cyber Security Course - Learning Analytics Report

Rutuja Jagtap (230118658)

2023-11-16

INTRODUCTION

This report aims to analyze trends, patterns, and anomalies across seven runs of a cybersecurity online course created by Newcastle University, using one of the most popular data mining techniques called the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework.

ROUND 1 of CRISP-DM

1. BUSINESS UNDERSTANDING

Business Background

Newcastle University made a massive open online course (MOOC) titled 'Cyber Security: Safety At Home, Online, and in Life.' This course is available to a global audience through FutureLearn, one of the largest British digital education platforms.

The course structure spans three weeks and comprises three sections, as follows:

- Section 1 : Exploring personal privacy online
- Section 2 : Online payment security
- Section 3 : Security in the future home

Every section includes a variety of resources, including discussions, articles, videos, and quizzes. The number of steps vary slightly in some runs of course, but the core content remains consistent throughout. Typically, there are approximately 18-19 steps in Section 1, 21-23 steps in Section 2, and 20-21 steps in Section 3 across the seven runs.

FutureLearn has provided data from their system for this learning analysis project. They anticipate receiving valuable insights from the analysis to inform future decisions regarding course design, enhance the learning process, and improve student engagement. The goal is to boost the course's popularity and overall success.

Business Objective

Business aims to understand the trends , patterns and gather valuable information from futurelearn data across seven course runs. In this round of CRSIP-DM, I am going to do analysis to determine the popularity of course in terms of enrollments happened in each run, I'm further going to show the learning outcomes of learners who participated in course quizzes and also the student engagement based on activities they studied in the course. These insights are valuable for stakeholders in measuring the overall effectiveness of the course. Later, based on this report they can analyze and formulate plans to enhance students' engagement and update course structure and process.

Success Criteria

The analysis produced must be effective enough for the stakeholders to consider and make sensitive business decisions. It is highly important to perform detailed study and transformations to the extreme raw data to get valid outcomes. Additionally, staying vigilant to biased data and not getting distracted from primary objectives. Segregating a big objective into small tasks and working on it will reduce confusion and enhance the quality of the report.

1.2 Assess Situation

Inventory Resources

Personnel:

```
- Stakeholder : Online Educator - Newcastle University
- Data Sponsor : Digital Platform - FutureLearn
- Analyst : Rutuja Jagtap
```

Data :

```
- Provider - FutureLearn
- Format - Zip folder named as MOOC FutureLearn dataset which contains
  online course step overview html document for all 7 run, 6-8 csv files of each run containing
  different kinds of raw data and course content overview file of all 7 runs in html document format.
```

Software:

```
- RStudio Version 2023.09.1+494 "Desert Sunflower"
- Microsoft Excel \
- Git version 2.42.0.windows.2
```

Requirements, assumptions and constraints

Latest dataset is required from FutureLearn. However, they currently have provided data from the year 2016 to 2018 and this reporting will be based on the provided data itself. Assuming course content and steps in the course are the same for all 7 runs as number of steps are a few here and there but main content is distributed similarly for the runs. Missing files, uneven data and unknown entries in records might constrain from gaining in depth information. Since data in the quiz response table has multiple attempts of a single question.So it is quite complex to determine accuracy rate. We will assume to consider only the first attempt of each learner to a question in order to define the percentage of accuracy rate per run.

Risks and contingencies

Due to old data the trend in data might have changed and finding of this analysis may not be valid for the current run of course. Moreover, there are serious issues with the quality of data provided to us. For example, data required for background check of learners is filled with more than 50% unknown values for most categorical variables which restricts us from gaining a lot of useful insights from the data.

Terminology

Successful Run - Run that stands out from other runs in positive context.

Costs and benefits

The insights from this report will assist stakeholders in making informed decisions for developing academic and business strategies. Furthermore, if the university decides to make improvements to the course for the benefit of learners, it will enhance their prospects in the future. Stakeholders can plan to offer exclusive course discounts to individuals from predominant categories or formulate promotional strategies if they feel they are running behind in terms of enrollments over the years.

1.3 Data mining goals and desired output

Business Goal - Identify the most successful run of course. Analyze the trend in enrollments, learning outcome and students activities over the duration of seven runs of course.

Data Goals - Identify right set of data to conduct analysis. Further, perform data wrangling wherever required.

Research question - Which was the most successful run? How did students perform over seven runs? What was the status of their activities?

Desired output - Brief comparison and evaluation between seven runs.

1.4 Produce project plan

Duration - The objective needs to be achieved within a fixed time period of 2 weeks.

Project Setup - R studio, Project Template, Rmarkdown file and Github setup are required. Futurelearn Data has to be loaded into the project data folder. Some useful libraries like dplyr, tidyverse, ggplot2, and readr are required to be installed as these are going to be extensively used during the analysis process. Report in rmd file format is the main file in which a report should be made and further knitted to pdf after completion of report.

Goals - Understanding business objectives and goals and noting down the prospects and hurdles in requirements, data and research methods.

Data preprocessing steps should be scripted in 01-A.R file which is located inside the munge folder.

Data - Understanding the data and importance of each variable. Checking and cleaning the data. Reading and transforming data to plot visualizations. Describing patterns and trends and trying to deep dive to details to find anomalies in data and state them in the report. Revisiting business objectives and working on data to find additional insights as needed. Highlighting difficulties faced in data collection or wrangling.

Analysis - using visualization techniques to describe findings.

Lastly, evaluating and sharing the research techniques used and results in the PDF format report to stakeholders.

2. DATA UNDERSTANDING

Data Collection

Futurelearn has already fetched raw data and provided it to us in a zipped folder for this investigation. It is extracted to the local system and placed in the project data folder to carry out seamless reading of files during the analysis and transformation process. Separate data are collected for all 7 runs which includes csv and pdf files.

Run - 1 dataset files (Total no.of files - 6):-

“cyber-security-1_archetype-survey-responses.csv”, “cyber-security-1_enrolments.csv”, “cyber-security-1_leaving-survey-responses.csv”, “cyber-security-1_question-response.csv”, “cyber-security-1_step-activity.csv”, ” cyber-security-1_weekly-sentiment-survey-responses.csv”

Run - 2 dataset files (Total no.of files - 7):-

“cyber-security-2_archetype-survey-responses.csv”, “cyber-security-2_enrolments.csv”, “cyber-security-2_leaving-survey-responses.csv”, “cyber-security-2_question-response.csv”, “cyber-security-2_step-activity.csv”, “cyber-security-2_team-members” ,cyber-security-2_weekly-sentiment-survey-responses.csv”

Run - 3 dataset files (Total no.of files - 8):-

“cyber-security-3_archetype-survey-responses.csv”, “cyber-security-3_enrolments.csv”, “cyber-security-3_leaving-survey-responses.csv”, “cyber-security-3_question-response.csv”, “cyber-security-3_step-activity.csv”, “cyber-security-3_team-members” ,”cyber-security-3_video-stats”, “cyber-security-3_weekly-sentiment-survey-responses.csv”

Run - 4 dataset files (Total no.of files - 8):-

“cyber-security-4_archetype-survey-responses.csv”, “cyber-security-4_enrolments.csv”, “cyber-security-4_leaving-survey-responses.csv”, “cyber-security-4_question-response.csv”, “cyber-security-4_step-activity.csv”, “cyber-security-4_team-members” ,”cyber-security-4_video-stats”, “cyber-security-4_weekly-sentiment-survey-responses.csv”

Run - 5 dataset files (Total no.of files - 8):-

“cyber-security-5_archetype-survey-responses.csv”, “cyber-security-5_enrolments.csv”, “cyber-security-5_leaving-survey-responses.csv”, “cyber-security-5_question-response.csv”, “cyber-security-5_step-activity.csv”, “cyber-security-5_team-members” ,”cyber-security-5_video-stats”, “cyber-security-5_weekly-sentiment-survey-responses.csv”

Run - 6 dataset files (Total no.of files - 8):-

“cyber-security-6_archetype-survey-responses.csv”, “cyber-security-6_enrolments.csv”, “cyber-security-6_leaving-survey-responses.csv”, “cyber-security-6_question-response.csv”, “cyber-security-6_step-activity.csv”, “cyber-security-6_team-members” ,”cyber-security-6_video-stats”, “cyber-security-6_weekly-sentiment-survey-responses.csv”

Run - 7 dataset files (Total no.of files - 8):-

“cyber-security-7_archetype-survey-responses.csv”, “cyber-security-7_enrolments.csv”, “cyber-security-7_leaving-survey-responses.csv”, “cyber-security-7_question-response.csv”, “cyber-security-7_step-activity.csv”, “cyber-security-7_team-members” ,”cyber-security-7_video-stats”, “cyber-security-7_weekly-sentiment-survey-responses.csv”

I have observed that run-1 has team-member and video statistics dataset missing and run-2 has video statistic file missing. Run-1 to Run-4 have weekly sentiment survey files but only columns are provided whereas column values are completely empty, run-5 has only one row entry. The Archetype file is empty for run-1 and run-2. Leaving survey response file is empty for run-1.run-2 and run-3.

Data collected consists of entries from the year 2016 to 2018.

Apart from csv files there are some html course overviews which display stepwise contents, quizzes and video materials. Separate overview files are given for different runs. There are on average 19 steps in the first week of course, 23 steps in second week of course and 20 steps in third week of course. However, in run -1, run-2 and run-3, there are some additional steps which were later amended in course structure but the number and topic of quiz step remains the same for all runs.

Data is quite raw and incomplete to draw confident conclusions hence next time Futurelearn should fetch complete data files.

Initially I used Microsoft Excel software to get a high level understanding of data. I also went ahead to use filter and min/max/avg features and noted the details of datasets.

Note: Modifications are not made directly to the MOOC dataset. Dataframes for the required files are created to read the dataset, and transformations are performed on these dataframes without altering the original data.

Data Description

Data collected as shown above includes information about learners' characteristics (archetype), employment status, educational background, age, country, gender, enrollment status, course video statistics, step activity, course feedback, learning progress and so on. To work on the business objective of this CRISP cycle i.e to study popularity or enrollment trends, learner outcomes and student engagement over seven runs, we require main files like enrolments.csv, question-response.csv and stepactivity.csv respectively.

1. To identify trends enrollments , I will be considering enrolled_at and learner_id columns from all the seven runs. As name suggests enrolled_at shows time at which particular learner enrolled for this course. It is a string type variable with format - YYYY-MM-DD HH:MM:SS in UTC timezone. Learner_id is a unique id given to a learner which is also in string type. Dataframe of run1, run2, run3, run4, run5, run6 and run7 are enrollment_1, enrollment_2, enrollment_4, enrollment_5, enrollment_6 and enrollment_7 respectively.
2. To analyze learner outcome , I will be selecting the columns :learner_id, correct and submitted_at columns for checking accuracy of multiple choice responses given by learners listed in quiz response files. Column:correct is of Boolean type and contains TRUE(for correct responses) and FALSE(for incorrect responses) while learner id and submitted_at are of string type.
3. To analyze student engagement, I will be considering columns: learner_id, step number and week number. All these variables are of categorical type. We want to see how many students visited each activity in each week from each run.

Data Quality :

- Size of the dataset is not the same for all runs. Size of data in run-1 is the largest of all.
- Couple of files are missing in some runs as mentioned in the data collection step of this CRISP round.
- Few step numbers are varying in initial runs. For instance, in Run-1, there are 18 steps for Section/Week 1, 21 steps for Section/Week 2, and another 21 steps for Section/Week 3. In Run-2 and Run-3, there are 19 steps for Section/Week 1, 23 steps for Section/Week 2, and another 21 steps for Section/Week 3. From Run-4 onwards till Run-7, there are 19 steps for Section/Week 1, 23 steps for Section/Week 2, and another 20 steps for Section/Week 3.
- Enrollment files have all unique entries, across all runs.
- Question responses dataset has multiple rows for the same learner id who has given multiple attempts to quiz. So it is quite complex to determine accuracy rate.
- Step activity dataset also contains multiple rows of data for a single learner as there are many steps in each sections.

Data Exploration

The main objective is to explore data across seven runs of the course. The three sub-tasks associated with this main objective are:

- Explore enrollment patterns
- Explore learning outcomes
- Explore learners' activity

1. Exploration of enrollment pattern

The enrollment count is crucial for analyzing the financial success achieved through the course. This analysis can indicate how many individuals have shown sincere interest in the course subject and have enrolled. I wanted to gain an overview of the strength of learners in each run of the course concerning the year in which enrollments occurred.

The clustered bar chart in Figure-1 illustrates that the maximum enrollments happened in the first run of the course, noted in the year 2016, reaching almost 14,394 enrollments. The second-highest enrollments occurred in the second run, starting from the end of 2016 and continuing until the beginning of 2017, with over 6,488 enrollments. There is a drastic fall in the number of enrollments in the second run.

Enrollments in the third, fourth, fifth, and sixth runs were 3,361, 3,992, 3,544, and 3,175, respectively, showing a slight increase in the fourth run and a significant drop from the fifth run onwards. Enrollments in the third run occurred entirely in 2017, while enrollments in the fourth and fifth runs happened in late 2017 and lasted until early 2018. On the other hand, enrollments in the sixth run occurred only in 2018. The least enrollments happened in the seventh run, with a count of 2,342.

I found that enrollments peaked in early 2016, dropped to a significant count in 2017, and gradually decreased to least numbers in 2018.

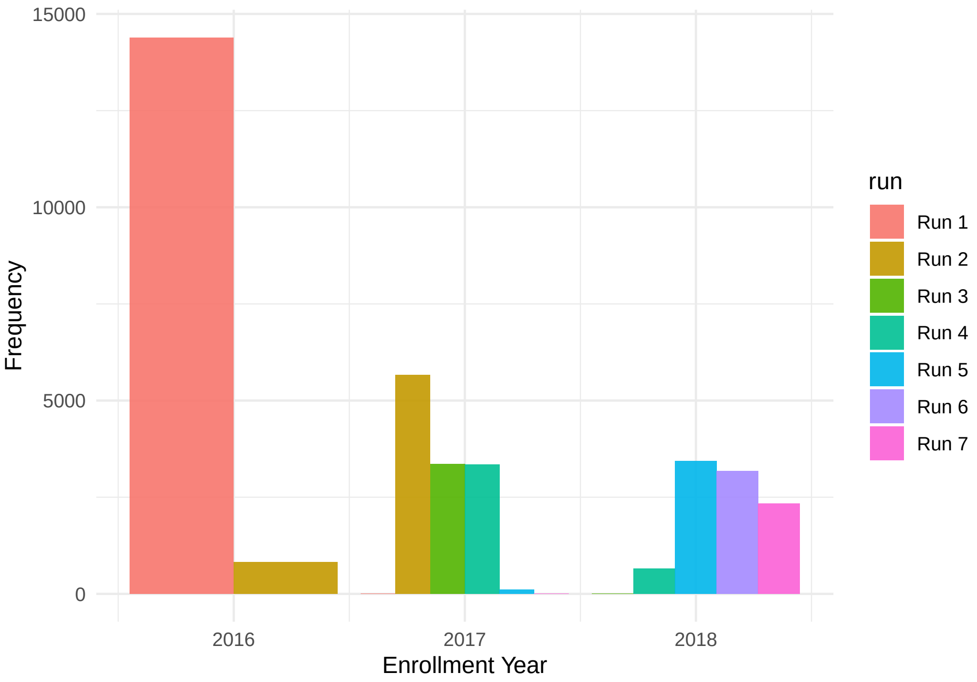

Figure 1: Barchart for Enrollments

Let's delve into the specific reasons why the enrollment for a particular run is not occurring within a single year. I have tried to illustrate this using a density plot in Figure-2 for better clarity, especially in terms of months. The density curve illustrates the spread of values in the 'enrolled_at' column over seven runs. Enrollment dates overlap from July 2017 onwards. A significant gap is observed between enrollments in run-2 and run-3. On the other hand, the maximum density of enrollments is noted in February 2018 in run-7. Meanwhile, enrollment dates spread widely in run-1 and run-2 over almost a year's time.

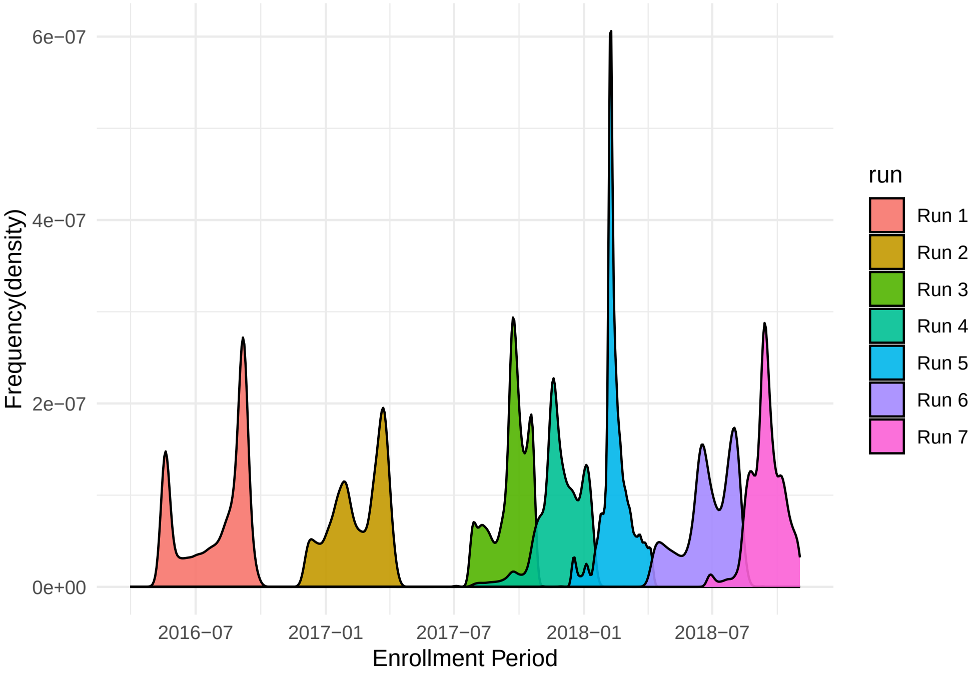

Figure 2: Density curve for Enrollments

To investigate deeper into the analysis of outliers, I decided to utilize a boxplot. The boxplot in Figure-3 revealed intriguing insights, indicating that Run-1 too includes some outliers that overlap with Run-2 and extend until late 2017. Although their number is limited but they exist and were not prominently visible in the previously plotted bar chart or density curve. In particular, the boxplot showcased that Runs 4, 5, and 7 exhibit a distinct clustering of outliers, forming a noticeable thick line. This observation implies that these specific runs are characterized by a higher concentration of outliers compared to the other runs, providing a discriminating perspective on the distribution of data points.

Another notable finding in this boxplot is that, in Run 1, the majority of data is concentrated in the lower quartile. In contrast, Run 5 lacks a distinct lower quartile, overlapping instead with the median of the plot

Data quality and preparation -

I chose to demonstrate analysis on enrollment data as it contains a lot of information and shows actual financial success through purchases made for enrollment.

The raw data includes an "enrolled_at" column in string format. Using the mutate function from the dplyr library, I transformed it into a more readable R format. Subsequently, I extracted only the year value and created a new column named "Enrollment Year." This "Enrollment" column was added to the dataframe for enrollment data file 1, denoted as "enrollment_1." Similar transformations were applied to the dataframes of enrollment data for the other runs as well.

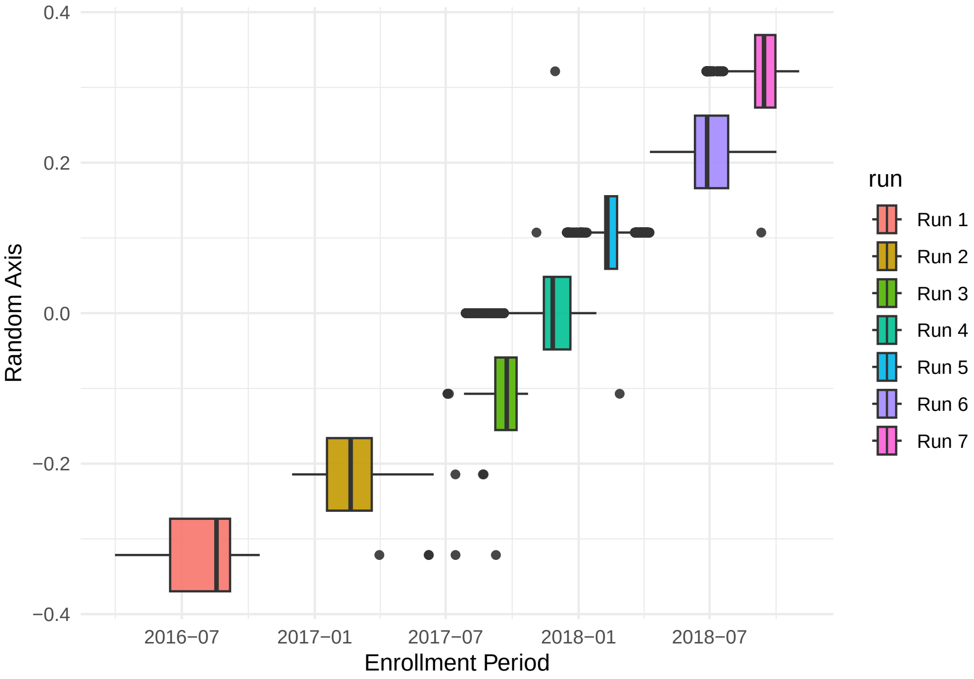

Figure 3: Boxplot for Enrollments

2. Exploration of learning outcomes

Detailed analysis on how learners are performing post-enrollment serves as a crucial criterion for determining the academic excellence of learners. It shall allow stakeholder to assess whether learners comprehend the topics covered in the course. To assess the knowledge of learners, quizzes have been incorporated into the course. I have considered this as an opportunity to extract insights and evaluate how learners have performed in these quizzes across all seven runs. The table below displays the percentage of correct and incorrect responses provided by learners in their first attempt for multiple-choice questions.

```
##                Correctness Run_1 Run_2 Run_3 Run_4 Run_5 Run_6 Run_7
## 1 Incorrect Response in %  43.9  41.2  41.7  43.8  41.4  43.1  42.5
## 2   Correct Response in %  56.1  58.8  58.3  56.2  58.6  56.9  57.5
```

It appears that learners in the second run have exhibited numerically strongest performance compared to participants in other runs. However, the overall performance across all runs falls within a relatively similar range of 56-58%.

Based on this observation, it can be concluded that while students have not achieved an excellent performance, there is room for improvement. The overall academic outcome across all runs appears to be average. This suggests an opportunity for targeted interventions or adjustments to enhance the learning experience and outcomes for the participants.

Data quality and preparation - I used the quiz response dataset for this analysis. I encountered some data issues when directly plotting the graph without any transformation. It yielded biased results, as the number of learners in run-1 is larger, leading to the perception that more correct responses were given by run-1 learners. Additionally, this dataset contains multiple attempts by most learners, making it challenging to determine the accuracy rate.

Therefore, I initially read the original data file into a dataframe and performed transformations on new dataframes. Firstly, I aimed at extracting only the first attempt to a question by grouping the data based on learner ID, quiz question, and question number. Subsequently, I selected the entry corresponding to the oldest response among all attempts. Lastly, computations were conducted on this data to generate a percentage summary matrix and further displayed in table as above.

I attempted to visualize the learning outcome trend through a graph. However, a similar pattern emerged due to negligible differences in the percentage of outcomes.Therefore, I used a table to illustrate exact percentage of correct and incorrect responses in each run. Additionally, using only true and false values without calculating percentages resulted in a biased outcome. For example, as there were more participants in run-1, it showed the highest number of true/correct responses for run-1. To address this, I opted to use percentages to present a more accurate and unbiased analysis of outcomes.

I chose quiz response data, as only this data of all the data files represented the questions attempted and correctness reported.

3. Explore learners' engagement

Understanding learners' engagement in the course is crucial to assess their commitment to completing each step. Stakeholders may be interested in knowing whether learners are genuinely benefiting from the content and actively participating in activities. This information can guide decisions on potential course amendments to enhance interactivity and engagement for future iterations of course. To assess this, I have presented plot to demonstrate the frequency of learners at each step week wise through a bar chart below for each run separately.

The following plots in Figure 4,5,6 and 7 illustrate the week numbers on the x-axis and the frequency of learners in each step on the y-axis. Each week is subdivided to represent step numbers, starting from 1.1, 1.2, 1.3, . . . , 1.7 for Week 1, 2.1, 2.2, 2.3, . . . , 2.21 for Week 2, and similarly for Week 3.

Run 1: - Out of 14,000 enrollments, 5,583 learners started the weekly activity by watching the welcome video.

Run 2: - From 6,000 enrollments, 2,666 learners watched the 1.1 activity.

Run 3: - Out of 3,361 enrollments, more than 2,300 learners started by watching the welcome video.

Run 4: - Out of 3,992 enrollments, 2,891 learners started watching videos of week-1.

Run 5: - Out of 3,544 enrollments, 2,671 learners started watching content.

Run 6: - Out of 3,175 enrollments, 2,244 learners started off with course.

Run 7: - Out of 2,342 enrollments, 1,595 learners showed engagement at beginning of course.

Throughout the three-week period, the number of learners visiting steps exhibited a gradual decrease, reaching the lowest count in each run. In run-1, despite a significant number of enrollments, only approximately one-third of the population engaged with the course by watching the welcome video. An interesting observation I found is that from run-2 onwards, there is a dip in the count at either the 4th last or 3rd last step in week-3 across all runs. Upon further investigation in the overview file for all runs, I observed that this step corresponds to step 3.18 'Test your understanding'. From run-2 onwards, many learners have not attempted this test, indicating a potential gap in their understanding of the course content. In contrast, in run-1, a considerable number of learners (approximately 1900) attempted the test, highlighting a difference in learner engagement and assessment participation.

Data quality and preparation - Date format was not in R readable format. Hence, that is changed by using dplyr function. The step_activity data has been meticulously filtered based on both week number and step number. Subsequently, the information is summarized and utilized to construct a comprehensive bar chart.

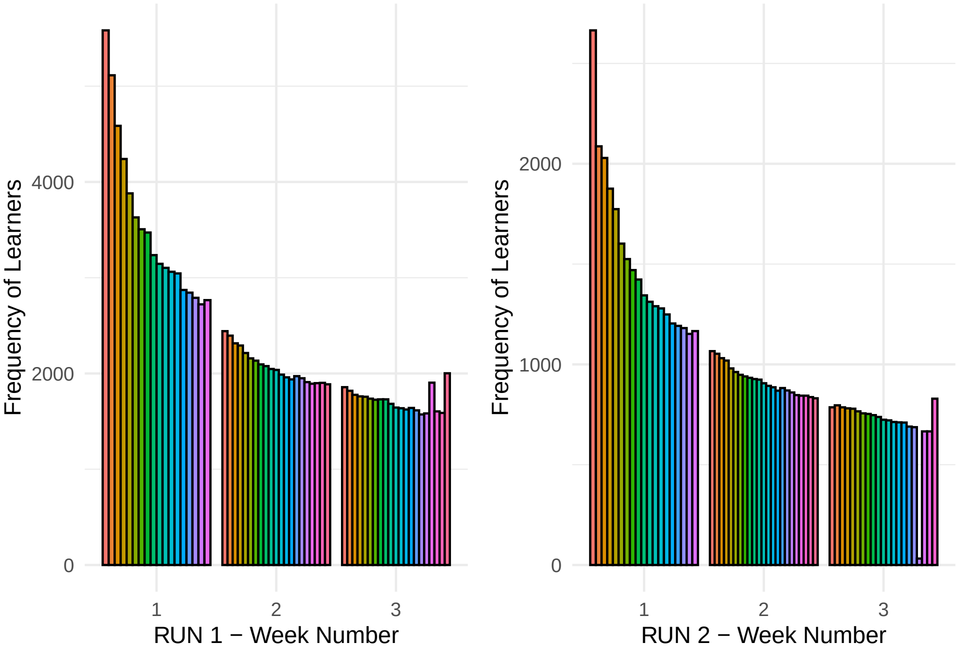

Figure 4: Learning activity chart of Run-1 and Run-2

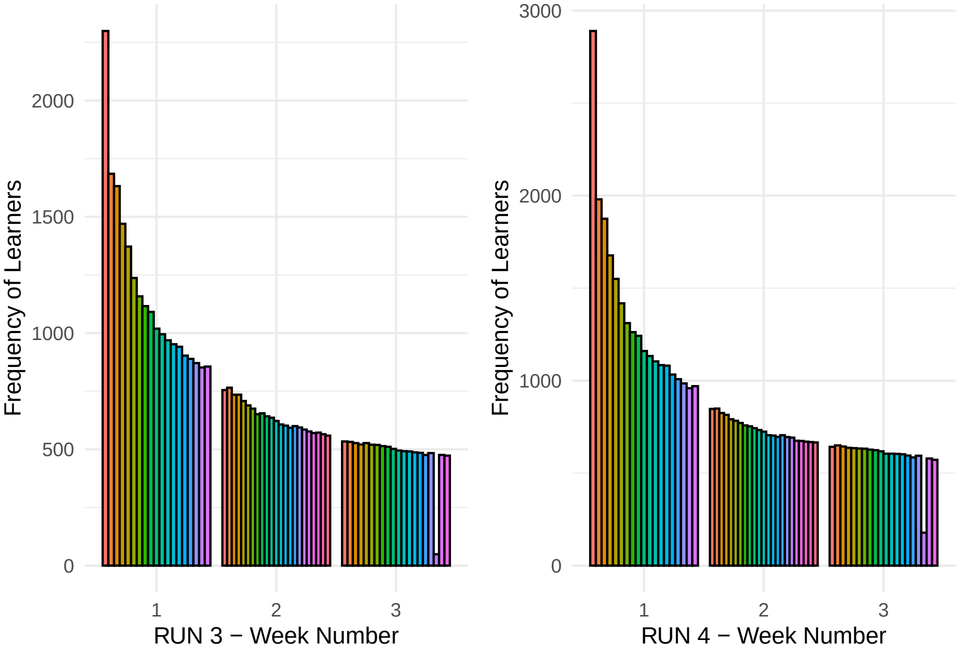

Figure 5: Learning activity chart of Run-3 and Run-4

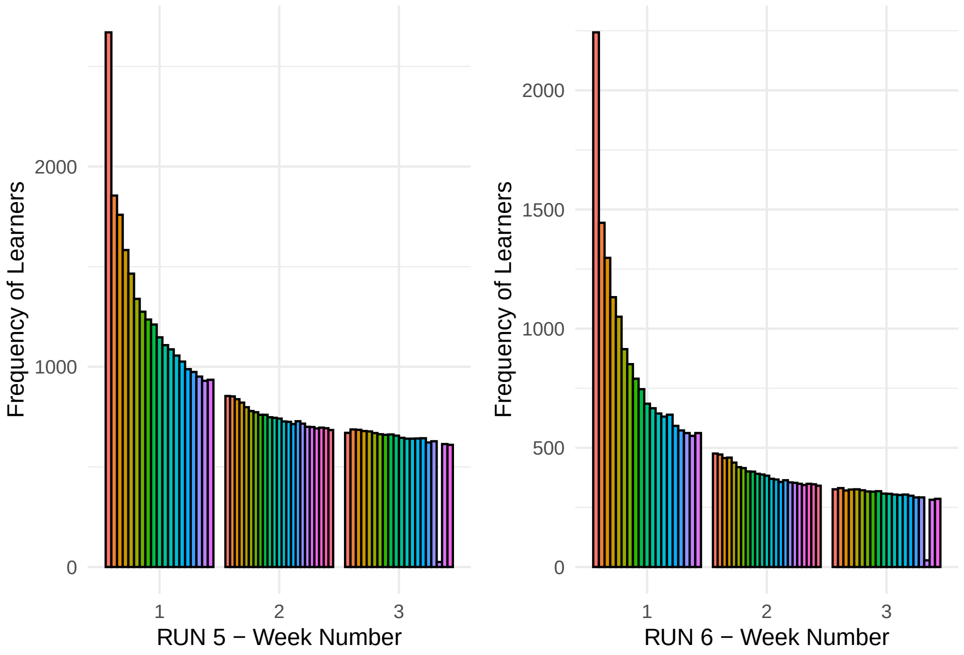

Figure 6: Learning activity chart of Run-5 and Run-6

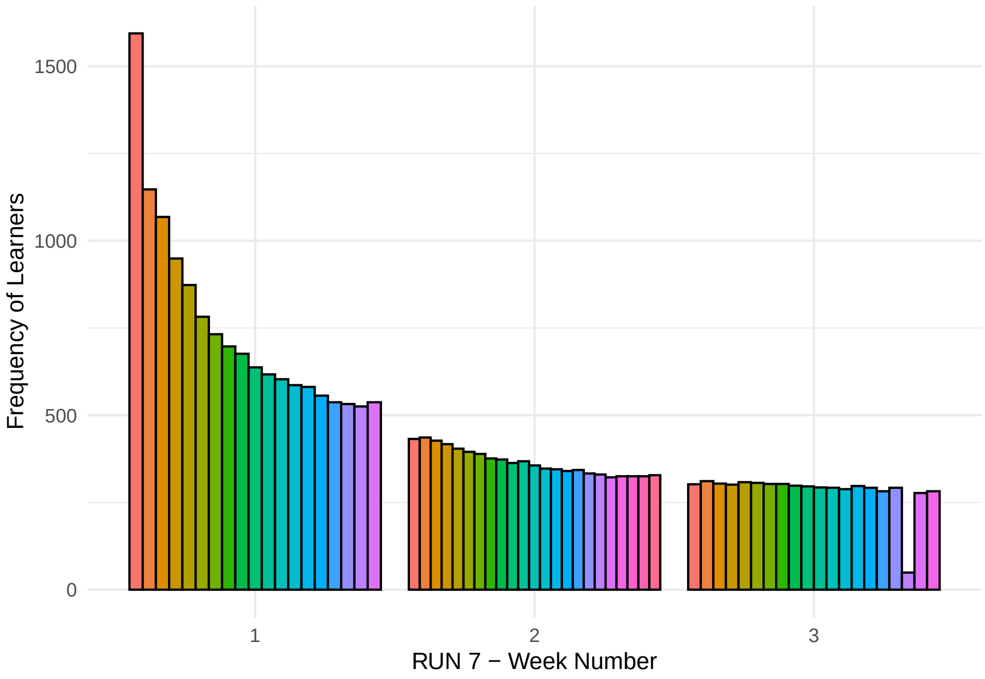

Figure 7: Learning activity chart of Run-7

3. EVALUATION OF CRISP-DM ROUND - 1

1. Among all the runs, Run 1 emerged as the most successful, boasting the highest enrollment numbers and overall popularity. This run has also proved to be most financially successful as many purchases of the course took place.The declining trend in enrollment over the years signals a crucial insight into the trajectory of business. This calls for urgent attention, prompting strategic planning to reverse the trend and set higher enrollment targets for the upcoming year.

2. Based on quiz responses, I attempted to assess the caliber of learners in each run to identify which run included high-performing students. However, I found that there was not a significant difference in their outcomes as. From the outcomes, it is clear that more than 50% of learners are able to provide correct answers, while less than 50% are giving incorrect responses.This analysis was found to reflect the sincere attempts of learners, even though they managed to provide correct responses in later multiple attempts.

3. The engagement in learning activities appears to be suboptimal, suggesting an opportunity to provide clear instructions encouraging learners to complete all steps for performance enhancement. I think the same trend is observed among learners in most online courses: learners are enthusiastic in the beginning but later start to lose interest and show less participation. Businesses need to make a plan to keep engaging learners through interactive content.

4. A potential reason for learners' less-than-excellent performance in the initial course quiz could be attributed to a lack of participation in step activities.

5. Run 1 stands out with the highest level of participation in step activities among all the runs.

6. The analysis conducted has successfully addressed all the outlined business objectives. The results obtained not only provide accurate data representation but also reflect a comprehensive understanding of the underlying factors. Throughout the analysis process, careful consideration was given to potential risks, constraints, and success criteria, ensuring a robust and insightful interpretation of the data. This approach enhances the reliability and relevance of the conclusions drawn, contributing to informed decision-making and strategic planning.

4. DEPLOYMENT

All the aforementioned findings and explorations will be compiled into a PDF format and presented to stakeholders as a comprehensive report. This document aims to provide stakeholders with valuable insights and information to facilitate decision-making based on the identified trends and outcomes.

ROUND 2 of CRISP-DM

1. BUSINESS OBJECTIVE

Building on the findings from Round 1, that Run-1 is the most popular and stands out in terms of number of enrollments and learners's participation compared to other runs. To gain further insights, in this round of crisp cycle, the business aims to know about the trends and patterns in background information of learners of Run-1 such as country, age, and employment status.

Research questions: Which country do most individuals in the course come from in this run? In which employment category and age range is the course most popular in this run?

2. DATA UNDERSTANDING AND PREPARATION

The dataset includes two country columns, namely “country” and “detected country.” After exploring data I found that the “country” column has over 80% unknown values, providing minimal information, I have opted to utilize the “detected country” column to determine the origin of individuals. For the “enrollment status” and “age gap” columns, there were many entries with unknown values. To handle this, I filtered out those entries and replaced them with NA values for further analysis.

3. DATA EXPLORATION

Part 1 :

In this section, I focused on the enrollment data from Run-1, organizing the countries based on their occurrence counts. The analysis involved identifying the most frequently detected country, the second most detected country, the third most detected country, as well as the least detected country. This approach provides a comprehensive understanding of the distribution of enrolled learners across different countries in Run-1.

As shown in Figure-8, the participation details from the data indicate that learners are predominantly from Great Britain (GB) with 34% strength , followed by India (IN) with 12% and United States(US) with 6% strength each and only one learner is from Burkina Faso (BF).Countries contributing less than 3% individually are collectively categorized as ‘Others,’ forming 48% of the total enrollment, demonstrating a diverse and distributed learner base as display in below piechart.

```
## Most Common Detected Country: GB    Count= 4939

## Second Common Detected Country: IN  Count= 1677

## Third Common Detected Country: US   Count= 865

## Least Common Detected Country: BF   Count= 1
```

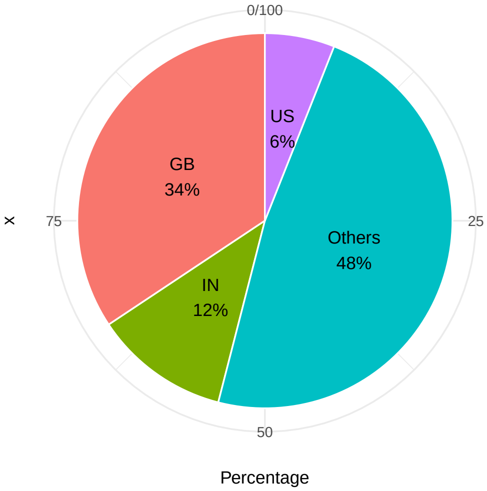

Figure 8: Demographic background of learners

Part 2 :

I have illustrated the distribution of learners across different age groups and employment statuses through the clustered bar chart in Figure-9.

1. **Age Distribution:**
   - A notable observation is the limited representation of learners below 18 years of age.
   - The highest concentration of learners is observed in the age group greater than 65, primarily consisting of retired individuals.
2. **Occupational Groups:**
   - The second-largest group comprises full-time workers in the age range of 26-65.
   - Following closely are full-time students falling within the 18-35 age bracket.
   - Apart from this, individuals who are in search of jobs are also showing considerable interest in the course. This can tell us that they want to expand their career in this domain and FutureLearn can suggest them with more of such similar courses in future.

Below visual representation provides a clear overview of the distribution of enrolled learners based on both age and employment status in Run-1.

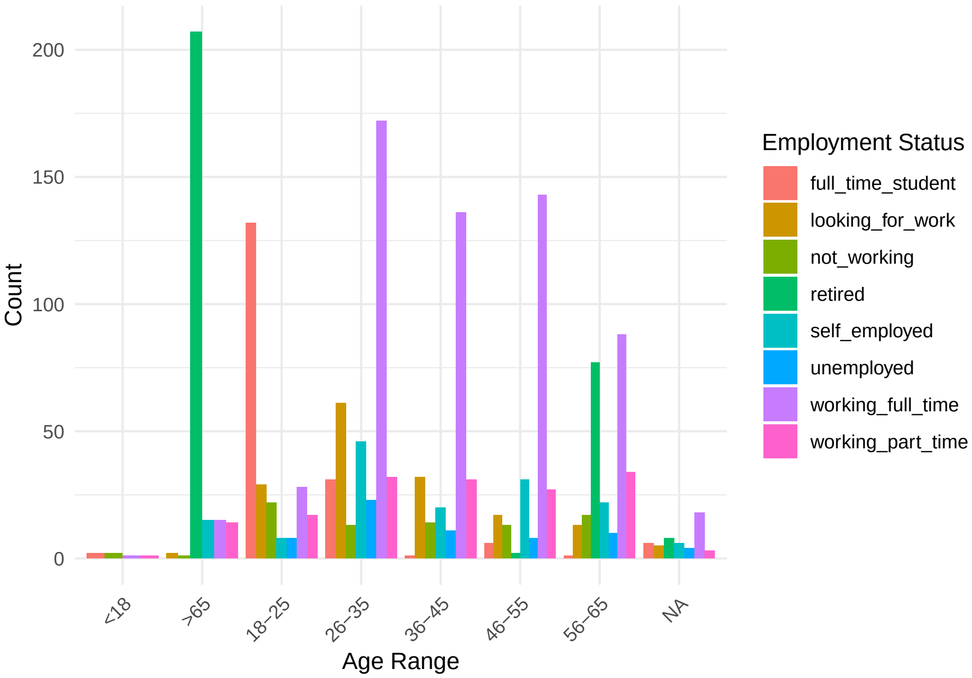

Figure 9: Age and Employment Status in Run-1

4. EVALUATION OF CRISP-DM ROUND - 2

1. In Run-1, the countries with the highest engagement in the cybersecurity course are Great Britain and India. Business can use this information to increase popularity of course in less participating countries by providing discounts or offers.
2. Individuals who are working full-time, retired individuals, and full-time students find the course most aligning to their interest.
3. Plots effectively capture trends, patterns, and anomalies.

The business objectives have been successfully achieved, and concise findings have been documented for stakeholders to make informed decisions.

5. DEPLOYMENT

Combined findings of CRISP DM round 1 and 2 will be kitted to pdf and given to Newcastle University and Futurelearn to look upon and take informed decisions.