Analysis Report

Pranav Sunil Raja

2024-11-14

```r
library(ProjectTemplate)
load.project()
```

```r
library(dplyr)
library(ggplot2)
```

Introduction

The objective of this project is to identify the key determinants of student success in the Massive Open Online Course (MOOC) titled "Cyber Security: Safety At Home, Online, and in Life," offered by Newcastle University through the FutureLearn platform. This study leverages raw data collected from seven different runs of the course, utilizing learning analytics techniques to uncover effective measures of student engagement. My analysis and findings will benefit Newcastle University's course instructors and individuals running similar online courses, as they seek to enhance student engagement rates and improve student performance. Employing the CRISP-DM framework, this report presents the results of two iterative cycles of data mining, which will serve as a valuable resource for improving the course's effectiveness and promoting it in a manner that aligns with the needs of the target audience.

Round 1

1. Business Understanding

This report is based on an inquiry into a Future Learn-hosted online course. Future Learn is an online learning platform that has worked with a number of world-renowned universities and organisations to provide a diverse set of courses. One of them is Newcastle University's Cyber Security, for which we will conduct a forensic examination in the form of a data analysis in this report. Overall this analysis might be beneficial to someone who is running the course or the course creator.

1.1 Background

Learning Analytics, a key facet of Data Science, involves the systematic measurement, analysis, and reporting of learner data to enhance educational out comes. Future Learn which is an educational website, they are primarily interested in anything that would improve and provide some actionable insights for informed decision-making, intervention strategies, and ongoing course refinement. We hope to use this method to assess the course's progress and determine where the course is prospering and where it is not. After we've addressed these difficulties, we'll be able to devise a strategy for improving the course as needed.

1.2 Business Objective

By employing the CRISP-DM methodology, this initiative unifies diverse data sources—ranging from on-campus facility usage to VLE interactions. The primary objective of this learning analytics project is to leverage data science methodologies for the FutureLearn MOOC Cybersecurity course. The goal is to measure, collect, analyze, and report data about learners and their contexts. Because Future Learn is an educational website, they are primarily interested in anything that would improve the learning experience, increase student interaction, and encourage students to enroll in more online courses. We hope to use this method to assess the course's progress and determine where the course is prospering and where it is not. After we've addressed these difficulties, we'll be able to devise a strategy for improving the course as needed.

1.3 Success Criteria

The success of this learning analytics initiative lies in the precise and accurate utilization of data science methodologies to enhance the FutureLearn MOOC Cybersecurity course. The results should be not only 1 relevant to stakeholders but also easily accessible and interpretable, fostering informed decision-making. The success criteria encompass ensuring data accuracy, relevance to stakeholders, accessibility and interpretability of results, applicability to learning enhancement, and the ability to inform strategic decisions for optimizing the online learning environment. Ultimately, success will be measured by the project's capacity to deliver actionable insights that contribute to the continuous improvement of the course and positively impact the learning experience.

1.4 Research Question

"Factors influencing completion, purchase and attrition of online course"

1.5 Inventory of Resources

Dataset: Data is made available from 7 runs of a massive open online course (MOOC) entitled "Cyber Security: Safety At Home, Online, and in Life". The 7 sets of raw data contains information on learners as they progressed through the course and some details on their profile. Course Overview: The seven course overview files describe course content and content type of each run.

1.7 Data Mining

The goal is to identify hidden patterns or trends within the data that are not immediately apparent. The detailed examination aims to uncover results to improve their marketing strategies, focus upon reducing the attrition rate and improve the purchase rates and overall resulting in the cost effectiveness and profit margin. By leveraging data-driven insights, the course analytics not only enhance the overall learning experience for students but also empower course designers and instructors to refine content, assessments, and interventions. The success is defined by the ability to unearth actionable knowledge and insights that can be translated into meaningful business strategies, improvements, or innovations.

1.8 Initial Assessment of tools and techniques

1. Git:
   Strengths: Git is an excellent version control system that enables collaboration, tracks changes, and provides a centralized repository for code and documentation.
   Opportunities: Ensure consistent and descriptive commit messages.

2. Project Template in R:
   Strengths: Utilizing a project template in R promotes organization, reproducibility, and efficient collaboration. It establishes a standardized structure for your project.
   Opportunities: Ensure that the project template aligns with best practices and is adaptable to changing project requirements.

3. ggplot, Tidyverse, dplyr:
   Strengths: These R packages, especially ggplot and dplyr from the Tidyverse, offer powerful tools for data visualization and manipulation. They support a clean and efficient coding style.
   Opportunities: Stay updated with the latest versions, explore advanced features, and consider incorporating additional Tidyverse packages as needed.

4. CRISP-DM (Cross-Industry Standard Process for Data Mining):
   Strengths: Following the CRISP-DM framework demonstrates a structured approach to data mining, ensuring that the project progresses through well-defined phases.
   Opportunities: Regularly review and adapt the CRISP-DM process to suit the evolving needs of the project. Emphasize clear communication and documentation.

2. Data Understanding

This phase encompasses the comprehensive documentation of data collection, description, and exploration. It involves providing information on the quality of the data, shedding light on its strengths, limitations, and overall reliability.

2.1 Data collection

Raw data is collected from the 7 course runs of a massive open online course (MOOC) entitled "Cyber Security: Safety At Home, Online, and in Life" made by Newcastle University. This data has information on characteristic information on every learner and steps of how they progressed through the course.

2.2 Data Files Description

For our cycle 1 analysis we only use enrollments. Following Data files can be viewed in the course runs.
1. Archetype Survey Response: This sheet contains the learner id's and their Archetypes.
2. Enrollments: The file consists of profile information of every learner id along with the enrollment and un-enrollment time frame.
3. Leaving Survey Response: This data sheet captures the leaving reasons and step activity of learners who unrolled from the course.
4. Question Response: This sheet contains the performance data of each student with the quiz content by storing their responses and attempts to solve each question. 5. Step Activity: This data sheet consists of the first visited and last completed time stamps of each activity that the learner's have visited. 6. Team Members: This file has details on the team members and their contribution to the course.
7. Video Stats: Video stats data sheet has information on the video content of the course. It has details like video duration, total views, downloads, views by percentage and total viewers on each device or demographic.
8. Weekly Sentiment Survey: The sheet captures comments and ratings of different learners for each week of the course.

2.3 Data Exploration

We want to understand different students profile who engage with the course. We're trying to find out if there are any patterns in the profiles across 7 runs. In my initial look at the data, there are some files that have information about students, but they don't help us with our main question. Some files, like Enrollments Data and Leaving Survey Response, have missing or unknown information. Also, not all files are available for every course run. For example, the archeotype survey responses and leaving survey response file is only there for the 4th, 5th, 6th and 7th runs of the course.

2.4 Data Quality

After exploring the data in detail, I identified potential data quality issues. This initial screening ensured no time was wasted on data sheets that are incomplete and do not meet the scope of our objective, providing an approach to verify the feasibility of its use and the ability to answer our defined question.
The enrollments data tells us when the user has fully participated, purchased statement, gender, country, age range and employment details in the course which would have been helpful to identify the demographics of the course but after inspecting further it is seen that it has many null or Unknown entries. So in order to get rid of these null or unknown values we ignore these entries in our demographic analysis . The incomplete data of Archetype survey response, Weekly sentiment survey response and Leaving survey response makes it difficult to draw exact insights from them. So in short we take only those which are completely available to do our analysis and ignore the null/NA values.

3. Data Preparation

Before moving forward to the next phase, I revisit the business understanding stage to incorporate insights gained during the data understanding phase. Specifically, I can provide additional comments regarding assumptions and constraints related to the chosen dataset and the exploratory analysis conducted. This ensures that our data mining process is aligned with a comprehensive understanding of the business context and any limitations associated with the data at hand.

3.1 Data Selection, Assumptions and Constraints.

For this phase of the analysis, I have chosen to focus on the enrollments file of all the 7 runs. This particular file provides a detailed overview of the timelines for each learner, indicating when they enrolled, unrolled, fully completed, purchase statement, gender, employment status, age range, highest level of education and country throughout the course based on each run.

```r
glimpse(master_extracted_data)
```

```
## Rows: 37,296
## Columns: 11
## $ run_cycle               <chr> "run1", "run1", "run1", "run1", "run1", "run1"~
## $ learner_id              <chr> "160d6600-ea0e-4568-bfa9-5d7cd5b8e61b", "4dc22~
## $ enrolled_at             <chr> "2016-08-10 14:28:49 UTC", "2016-05-24 17:34:3~
## $ fully_participated_at   <chr> "", "", "2016-09-22 16:56:03 UTC", "", "", "20~
## $ purchased_statement_at  <chr> "", "", "", "", "", "", "", "", "", "", "", ""~
## $ gender                  <chr> "Unknown", "male", "Unknown", "Unknown", "Unkn~
## $ age_range               <chr> "Unknown", "46-55", "Unknown", "Unknown", "Unk~
## $ highest_education_level <chr> "Unknown", "university_degree", "Unknown", "Un~
## $ employment_status       <chr> "Unknown", "working_part_time", "Unknown", "Un~
## $ employment_area         <chr> "Unknown", "teaching_and_education", "Unknown"~
## $ detected_country        <chr> "GB", "PE", "NG", "UG", "IM", "NO", "GB", "GB"~
```

Given the presence of null values in most of the columns, we operate under the assumption that these null values are not influencing or affecting the final Leveraging this assumption. Upon having look into data availability during the second stage, it's evident that incomplete data presents a constraint. Certain data files are not consistently present in every course run, introducing variability across runs and complicating a comprehensive analysis of the entire dataset. Given the rogue data, the majority of my analysis will be focused on removing or ignoring the Unknowns/NA across various runs of the course.

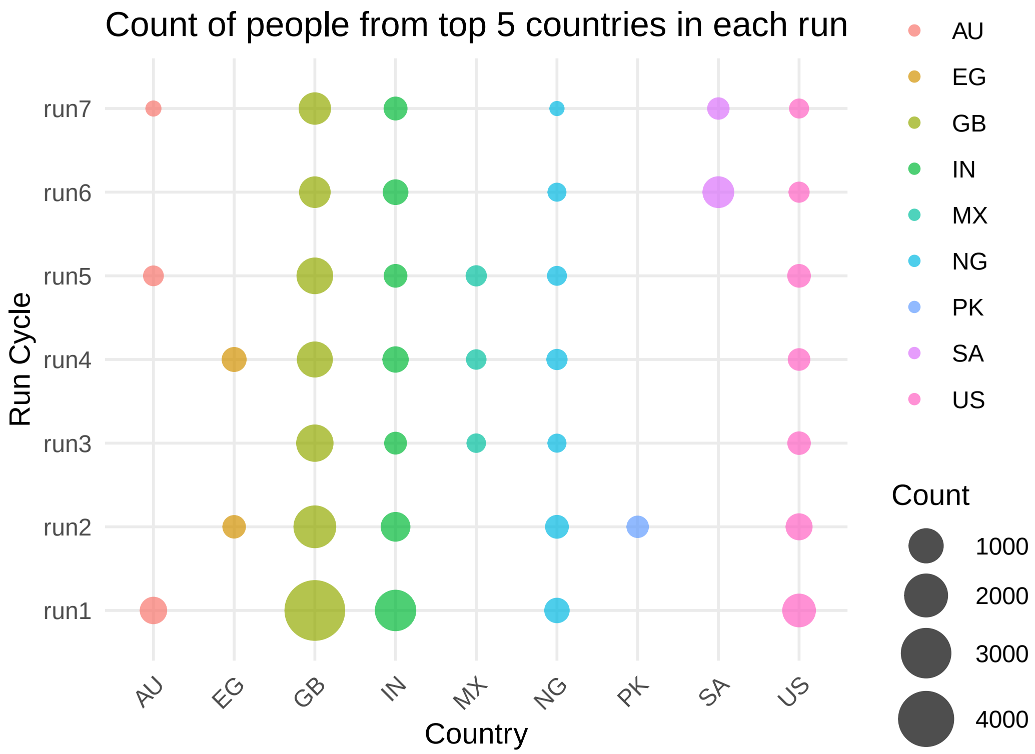
3.2 Data Construction and Formatting

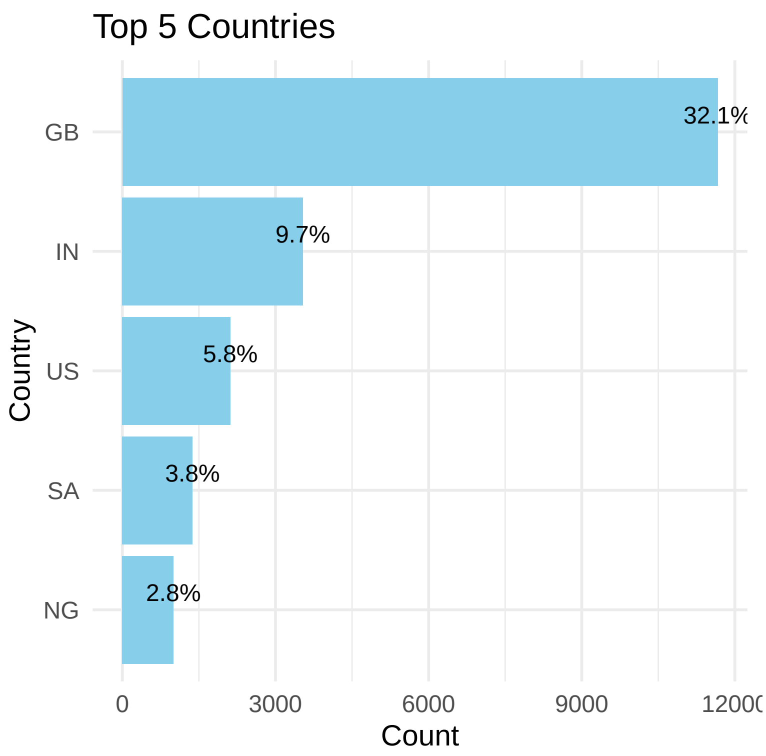
The dataset contains profile data for 37296 learners across 7 runs. A master dataset is constructed combining all seven runs and is worked on that. All the Unknowns and NA values are filtered out using dplyr functions so that consistency is maintained and we have also observed that these have no effect hence are filtered out. By calculating the demographics for each run, we understand the variation across each run and then compare with the demographics of overall total which contains all runs. run_cycle indicates the run number. The final result allows us to observe some trends.

4. Modelling.

Upon concluding the data preparation phase, I revisited the earlier stages to verify adherence to the initial plan and assess potential impacts on the upcoming analysis. Finding no deviations or issues, I proceeded to complete the exploratory data analysis to answer our question - "Can we identify the patterns of student enrollment in the course?"

4.1 Enrollment to country relationship

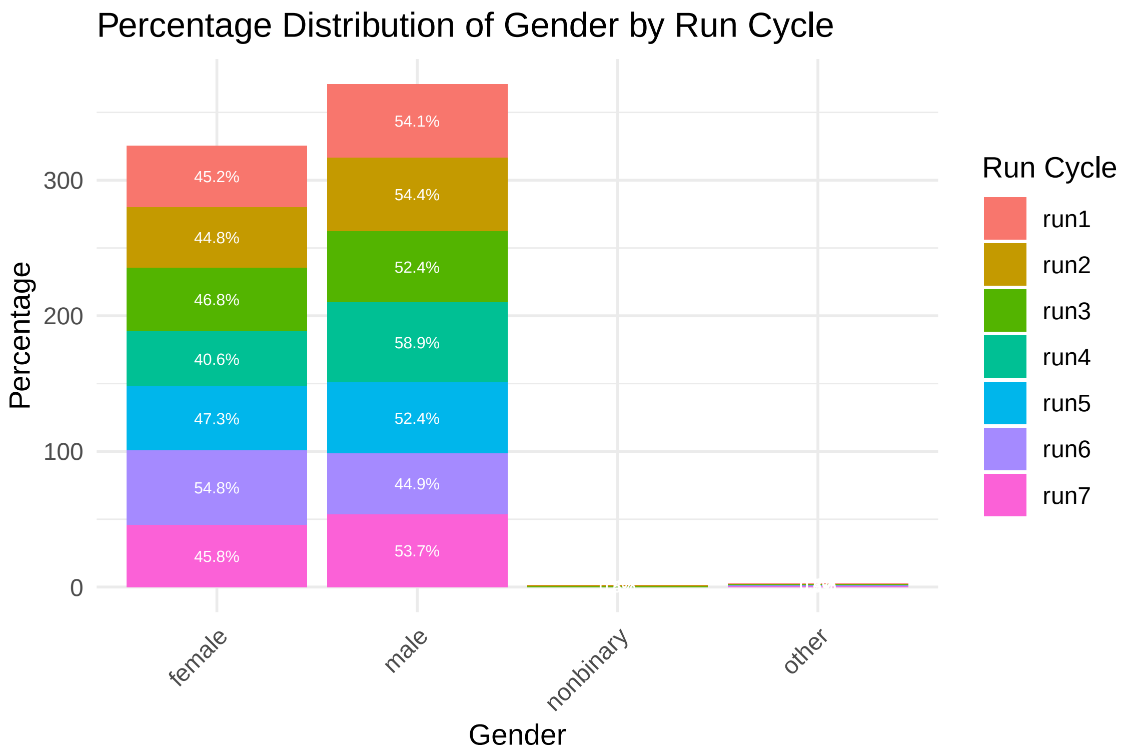


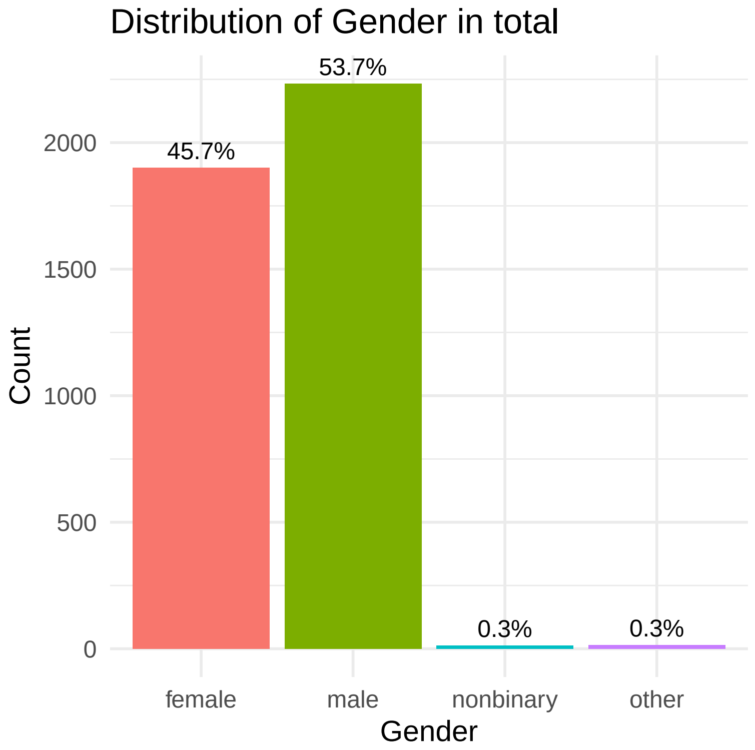


This analysis shows that the top three countries are GB, IN, and US. They have consistently been the largest contributors of learners, with GB maintaining a high position accounting for over 32% of the total. However, the relative share of these top countries has decreased over time, indicating a growing diversity in the geographic distribution of participants, as emerging markets like South Africa (SA) and Nigeria (NG) have seen their share of participants rise across the run cycles, while the. Though, in the earlier run cycles we have a high number of participants with an unspecified country of origin, traditional powerhouses have remained strong.

The varying participation patterns across countries highlight the need for the program organizers to develop targeted recruitment and retention strategies to optimize the program's appeal and accessibility for a more diverse global audience.

4.2 Enrollment to gender relationship



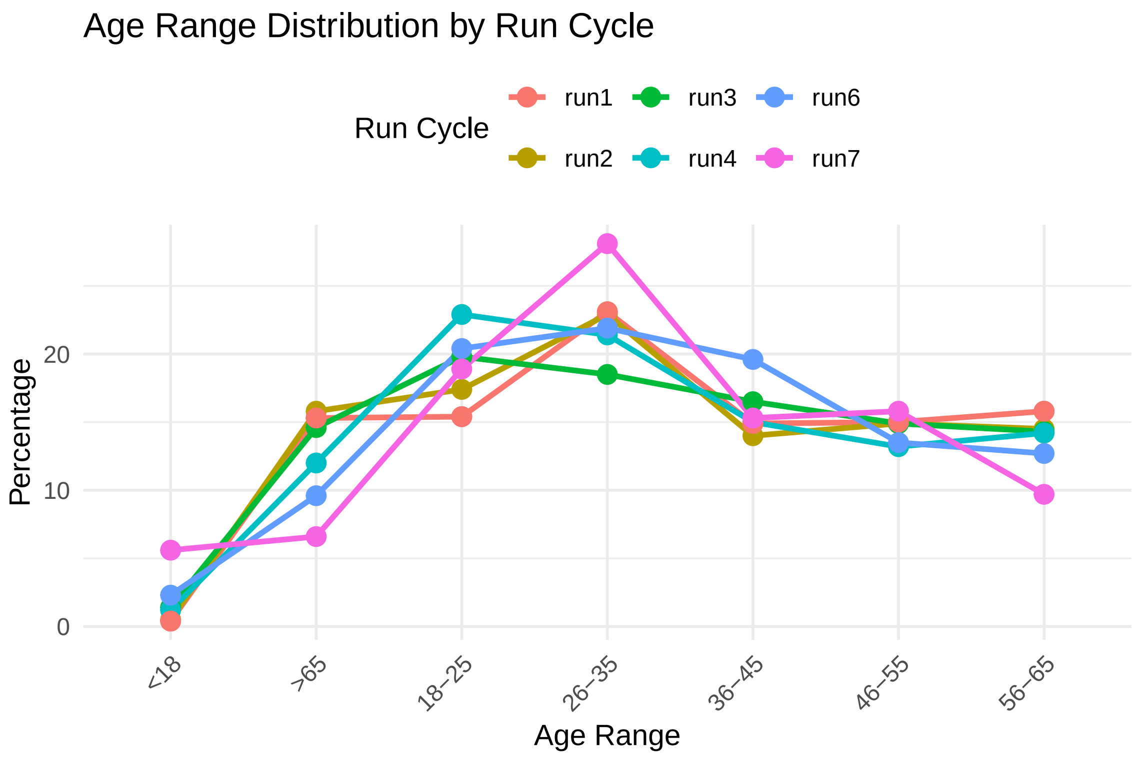


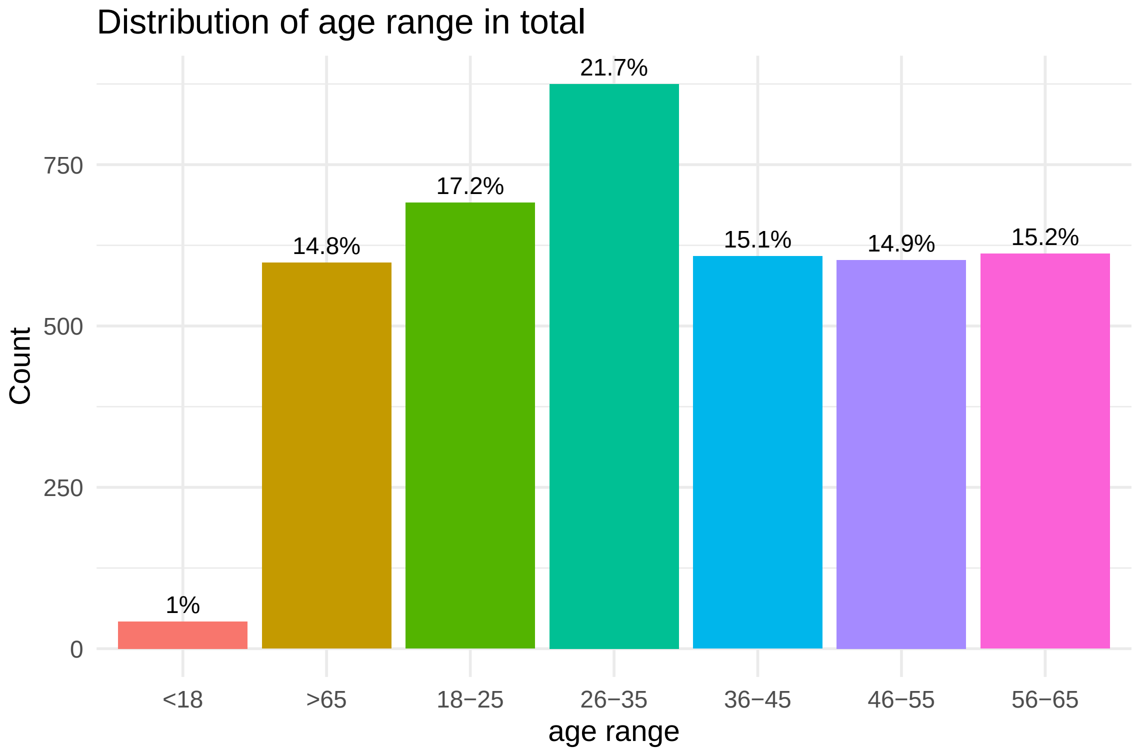
This analysis shows gender distribution of participants which has some fluctuations across the different run cycles. While male participants have always been consistently forming the majority, the proportion of female participants has seen a steady increase from 1st Run to &th Run. This suggests that the program is gradually becoming more appealing to a wider range of gender identities. However, the data also shows some variations, such as the decrease in the proportion of female participants from the third run to 4 fourth run but increases again in subsequent cycles.

In the total gender composition across all run cycles, male participants have the largest share of 54.0%, followed by female participants (46.0%). While the male majority is persistent, the gender gap appears to be narrowing, indicating a more balanced representation. Additionally, the data reveals a small and consistent presence of participants who identify as non-binary or other gender identities, accounting for 0.7% of the total.

These fluctuations highlight the need to closely monitor the gender dynamics and understand the factors driving these changes over time.

4.3 Enrollment to age range relationship



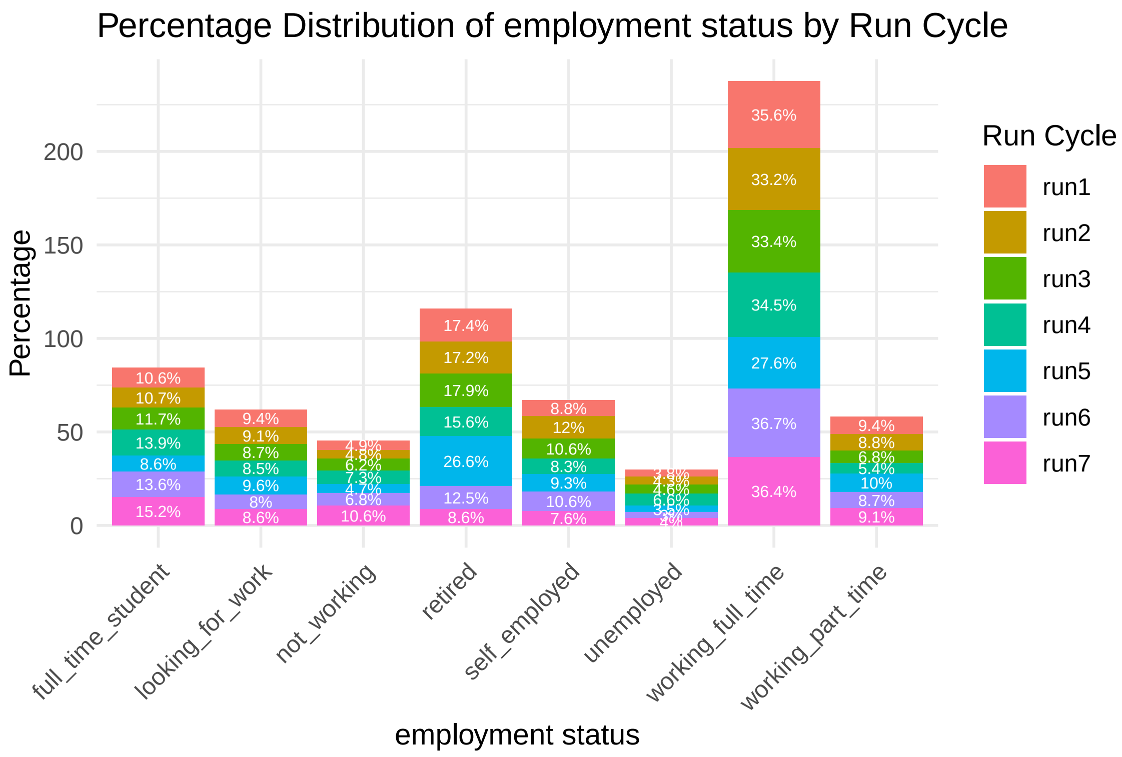


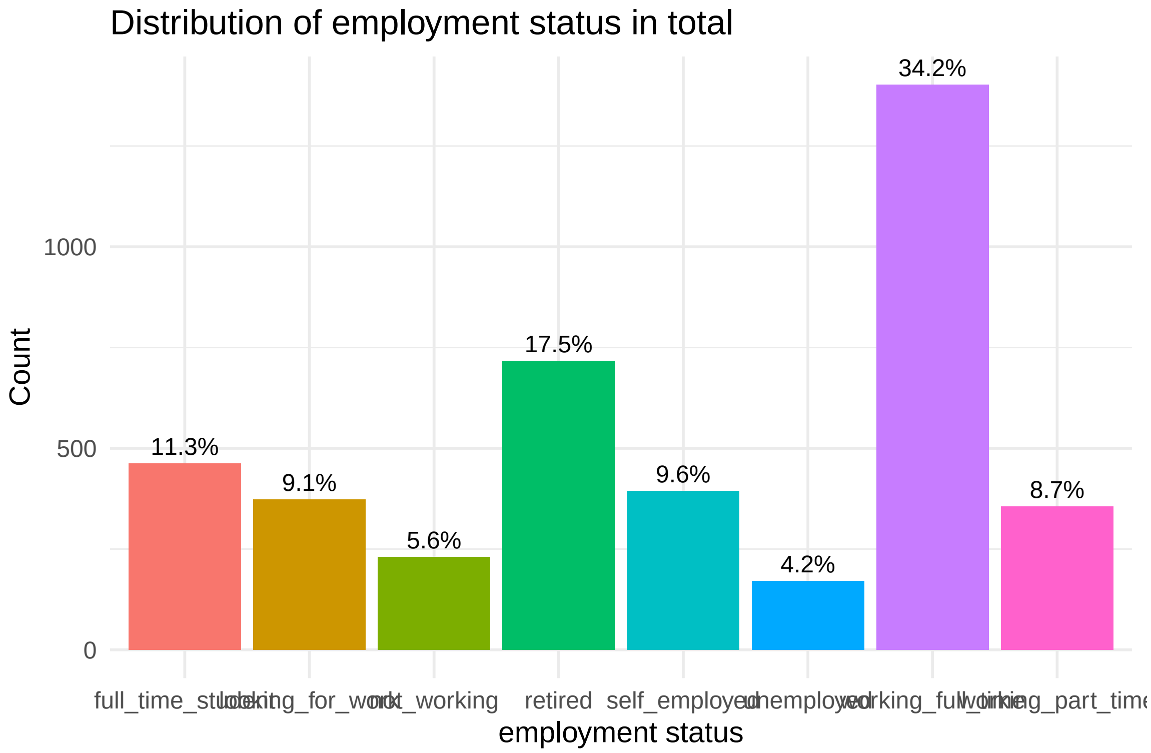
The analysis of age distribution of participants exhibits some fluctuations across the various run cycles. In the earlier run cycles, the 26-35 age group consistently formed the largest segment, accounting for around 23% of participants. However, this proportion decreased in the later run cycles, with the 18-25 age group becoming more prominent, increasing in the first run to the seventh run. Additionally, the representation of participants aged 56 and above has varied, with the 56-65 and over-65 age groups seeing both increases and decreases in their relative count across the different cycles. These shifting age dynamics suggest that the program's appeal may be evolving to attract a more diverse range of age groups over time.

Analyzing the total age composition across all run cycles, the data reveals that the 26-35 age group forms the largest segment, followed by the 18-25 and 56-65 age groups. The remaining age groups 36-45, 46-55, and over-65 have relatively similar shares. Interestingly, the youngest participants (under 18) account for only 1.0% of the total.

Interestingly, the youngest participants (under 18) account for only 1.0% of the total, indicating that the program may need to explore targeted outreach and messaging to engage a younger demographic. Understanding the unique needs and preferences of each age group can help the program organizers develop more tailored strategies to attract and retain a diverse participant base.

4.3 Enrollment to employment status relationship



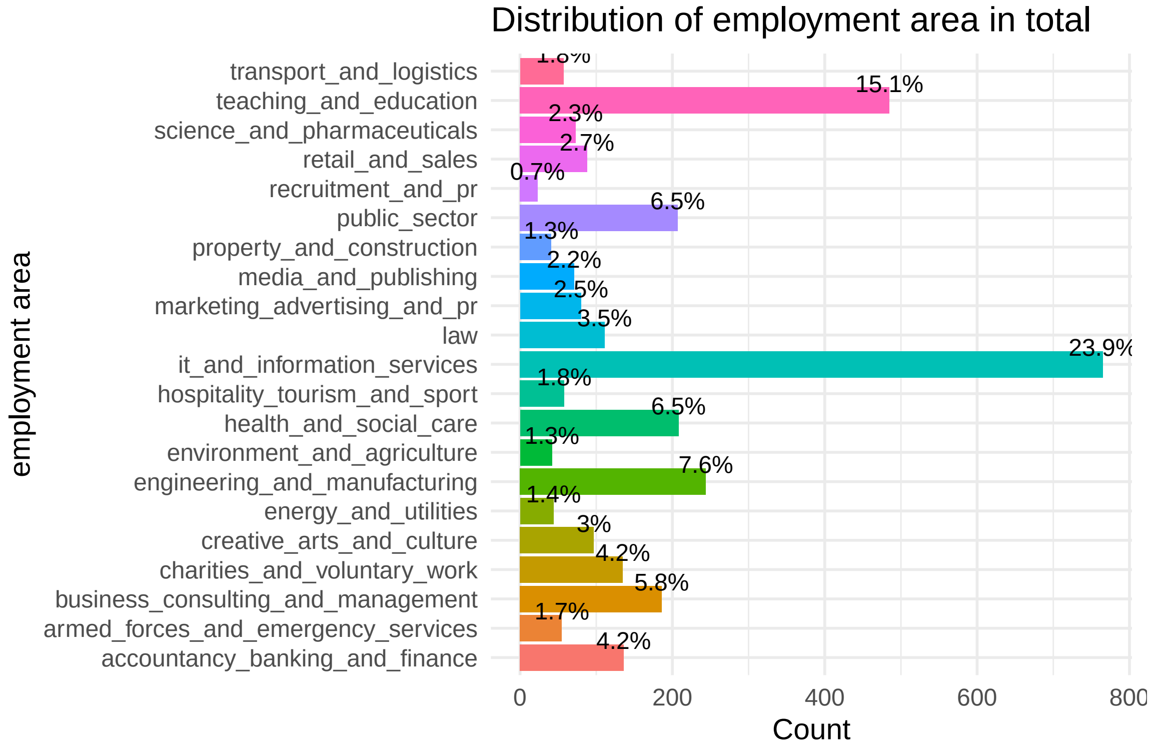


The analysis of distribution of participants across different employment statuses exhibits some fluctuations over the run cycles. In the earlier cycles, the “working full-time” category consistently formed the largest segment, accounting for around 35% of participants. However, this proportion decreased in the later run cycles, reaching to 36.4% in the latest run. Conversely, the “retired” category saw an increase, from the first run to the fifth run and then dropped in the seventh run. The representation of other employment statuses, such as “full-time student,” “self-employed,” and “looking for work,” also varied across the different run cycles. These shifting trends suggest that the program’s appeal may be evolving to attract participants from diverse employment backgrounds over time.

In the total employment status composition across all run cycles, the analysis reveals that the “working full-time” category forms the majority, followed by “retired” and “full-time student”. The remaining categories, including “self-employed,” “looking for work,” “working part-time,” “not working,” and “unemployed,” account for smaller but significant shares.

Understanding the unique needs and preferences of participants across these different employment statuses can help the program organizers develop more tailored strategies to attract and retain a diverse participant base. Identifying any barriers or challenges faced by specific employment groups can also inform the program’s design and support services to better accommodate their requirements.

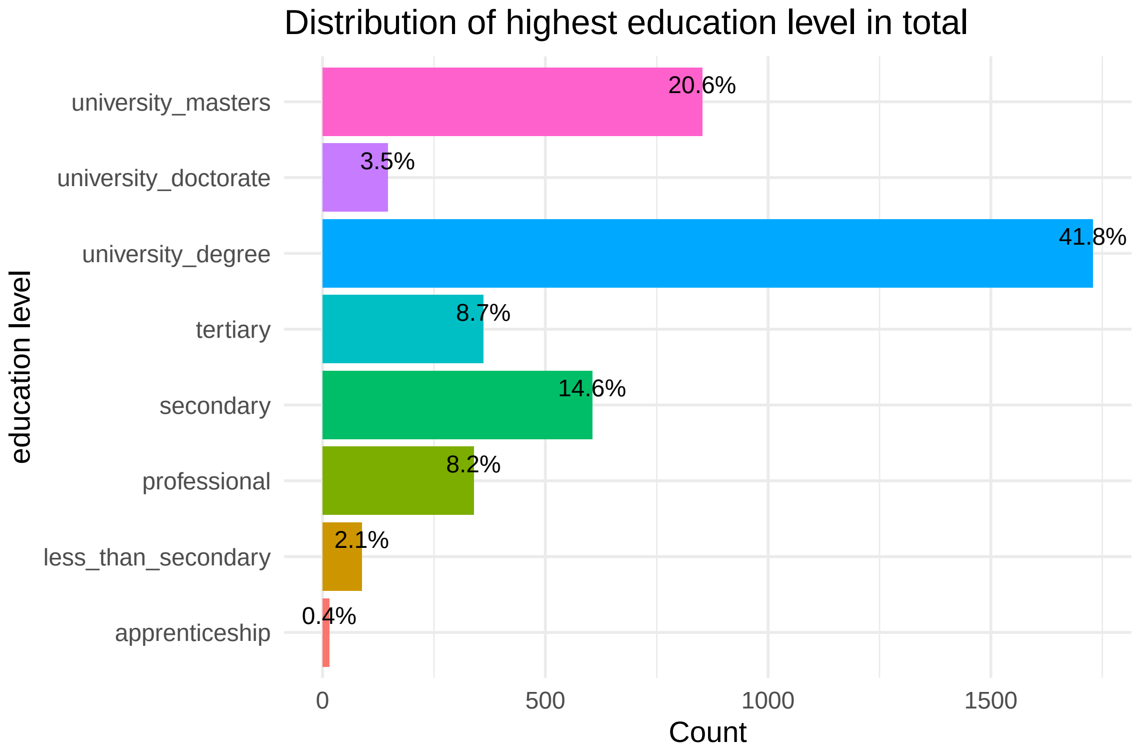
4.4 Enrollment to employment area relationship



The analysis reveals a diverse representation of employment backgrounds, with the program attracting a strong concentration of professionals from the IT and Information Services , Teaching and Education and Engineering and Manufacturing sectors. This suggests the program has particular appeal among individuals in technology, education, and technical fields. However, the data also shows a relatively broad distribution of participants across other industries, including Health and Social Care, Public Sector, Business Consulting, and Finance. The main sectors have the most people, but we could work on getting more participants from areas like Retail, Media, and Hospitality.

By understanding the unique needs and barriers faced by participants from diverse employment backgrounds, the program organizers can refine their strategies to ensure equitable access and maximize the program’s appeal across a wide range of professional sectors.

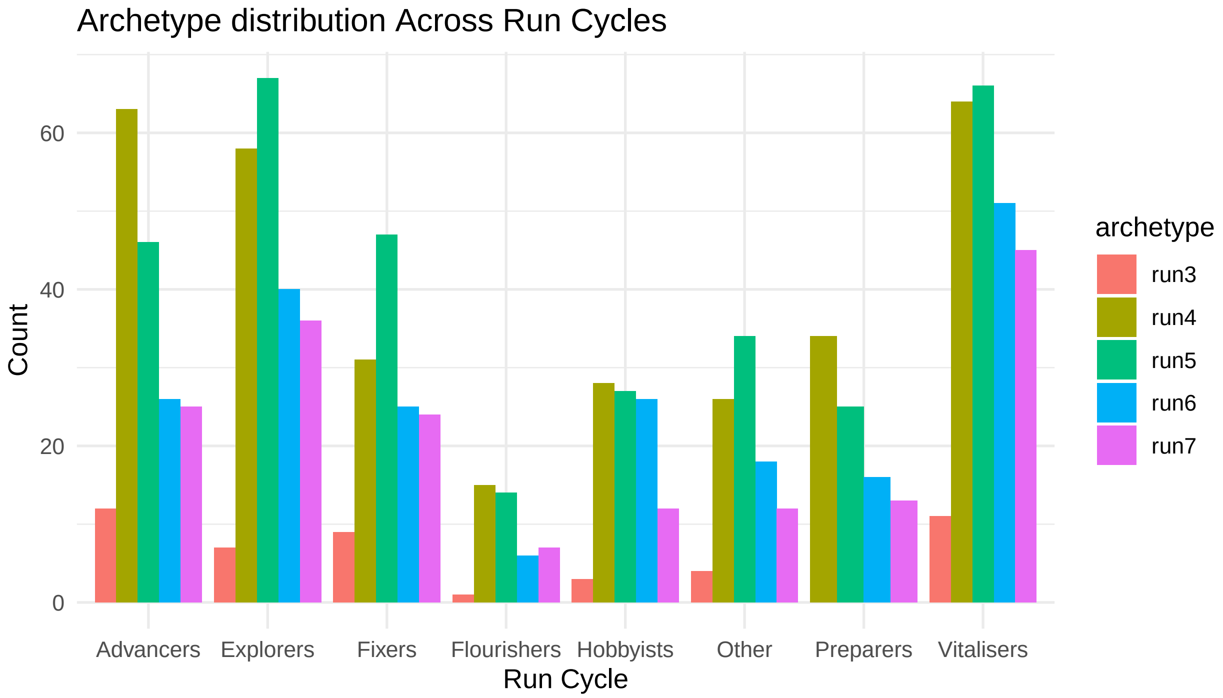
4.5 Enrollment to Highest education earned relationship



The participant data analysis reveals a highly educated audience, with the majority holding university degrees or master's degrees. This suggests the program is attracting a sophisticated, academically achieved group of learners. However, the representation something more aswell, highly credentialed, with significant portions of participants holding secondary, tertiary and professional qualifications as well as a small percentage with doctoral degrees and even less than secondary education. This indicates the program's ability to attract to a wide range of learners, from those with foundational education to those with specialized, advanced degree's.

Understanding the unique perspectives and learning needs of participants across this spectrum can help the organizers tailor the program content, delivery, and support services to optimize the learning experience for this diverse participant base.

4.6 Different types of Archetype in all runs



This analysis shows a mix of user archetypes, with Vitalisers being the largest group followed by Explorers and Advancers. Fixers, Hobbyists and Preparers make up smaller segments and the other category accounts for a very small number. This suggests that the course has appeal across different user motivations and behavior. From those focused on personal growth and discovery to those more interested in problem-solving and optimization. By understanding the unique needs and preferences of these user groups, development of tailored features, personalized messaging, and targeted engagement strategies to effectively prepare to the diverse customer base.

4.7 Paricipation Rate

```
participation_rate
```

```
##   total_enrolled total_participated percentage
## 1          37296               2154        5.8
```

This shows a participation rate of only 5.8%, with 2,154 participants out of a total enrollment of 37,296.

This relatively low participation rate suggests there may be a need of opportunities to examine barriers to participation and implement strategies to improve engagement. Some possibilities to explore evaluating the program content and format, and gathering feedback from both participants and non-participants and also identifying the key factors driving the current participation levels could inform initiatives to increase active involvement and maximize the program's reach and impact.

4.8 Purchase Rate

```
purchase_rate
```

```
##   total_enrolled total_purchased percentage
## 1          37296             289        0.8
```

The 0.8% purchase rate out of 37,296 total enrolled users indicates a significant need to analyze and optimize conversion factors. Key areas to focus on could include marketing, digital marketing, product positioning, user experience, and pricing in order to drive higher purchasing activity and maximize revenue potential. And also find ways to improve the value of their certification so that people find the need of purchase.

Evaluation

In this conclusive phase, the analysis of the FutureLearncybersecurity course data has successfully met the defined business objectives and success criteria. It has delivered valuable insights into student engagement patterns and areas for potential improvement. By exploring the demogrphics, a comprehensive understanding of learner behaviors has been achieved. The visualizations have proven effective in communicating these insights, although there's room for refinement in specific graphical representations. Despite challenges like incomplete data and variations across course runs, the overall quality of the dataset has been decent enough to capture general trends. Looking ahead, the plan is to conduct an in-depth analysis of student dropout ratios, aiming to gain further insights.

Round 2

1. Business Understanding

1.1 Business objectives and success criteria

In the upcoming second cycle of our CRISP-DM analysis, we intend to delve deeper into understanding student patterns by scrutinizing diverse learning materials in the cyberSecurity course. Our primary questions for this phase include investigating the correlation between student leaving reason with last completed step and the analysis of few other factors. Additionally, we aim to explore if the leaving reason and the steps or week have any correlation. This refined focus aims to uncover nuanced insights into the influence of various learning materials on student engagement, guiding efforts to optimize course content and enhance the overall learning experience.

The success of the second cycle of our analysis will be determined by the extent to which we uncover actionable insights about student un-enrollement patterns with different leaving reason. Success criteria include identifying clear correlations between them. Additionally, the ability to discern specific content types that significantly impact student dropout rates or sustained engagement will be crucial. The results should contribute valuable information for optimizing course content, addressing learner challenges, and ultimately enhancing the overall effectiveness of the cyberSecurity course. The analysis should be presented in a clear, accessible format, facilitating easy interpretation and utilization by stakeholders for informed decision-making.

1.2 Expenditure

1. Time and Resources: Conducting a comprehensive analysis requires a significant investment of time and resources, including data collection, preprocessing, and iterative exploration.
2. Data Quality Challenges: Dealing with incomplete or inconsistent data poses challenges that may require additional efforts for cleaning and validation.

3. Analytic Tools: Utilizing industry level analytical tools and software may incur costs, particularly if specialized software or computing resources are needed.
4. Expertise: Employing skilled analysts or data scientists to perform the analysis adds to the overall cost.

1.3 Benefits:

1. Actionable Insights: The analysis provides actionable insights into student engagement patterns, dropout tendencies, and potential areas for improvement in the cyberSecurity course.
2. Informed Decision-Making: Stakeholders can make informed decisions regarding course content optimization, addressing learner challenges, and enhancing overall course effectiveness.
3. Enhanced Learning Experience: Improving the course based on analysis results can lead to a more engaging and effective learning experience for students.
4. Strategic Course Development: Understanding correlations between different types of socio demographics, student engagement and leaving reasons can inform strategic course development for future offerings.

2. Data Understanding

After defining business objectives I moved onto the next phase of my CripDM cycle.

2.1 Data Description

In this analysis, we use a key file leaving response survey from most of the runs available. The leaving survey file contains insightful information such as the last step, last step number, last week number and the leaving reason. Key features in this file include leaving reason. Analyzing this data will aid in identifying reasons for student dropout due to specific reason and learning outcomes.

2.2 Data Quality

To ensure the integrity of the analysis, it was necessary to address null values in the columns of the leaving survey responses file. Rows with blank were removed to enhance the quality of the data and facilitate a more robust analysis. Despite this preprocessing step, the overall data quality is deemed satisfactory, providing a solid foundation for conducting thorough analyses and drawing meaningful conclusions from the available information.

3 Data Preparation

After conducting a comprehensive analysis of the data, I revisited the earlier phases to check for any additional considerations. Finding no further modifications needed, I transitioned to the subsequent phase, namely the data preparation stage of this cycle.

To enhance the clarity and visual representation of the graphs, several data formatting steps were implemented. In the last completed step & last completed week rows with null values in the column were removed. These formatting adjustments contribute to the overall readability and effectiveness of the subsequent visualizations.

4 Modelling

After formatting and cleaning the data properly I moved on to the most interesting part of the process, performing exploratory data analysis.

4.1 Purchase rate of people who have fully compleated the course

```
effective_rate
```

```
##   total_completed total_purchased percentage
## 1            2154             289       13.4
```

We see around 13.4% effective rate, with 289 purchases out of 2,154 total completions, indicates an opportunity to further optimize the conversion process. Maybe to boost certificate purchases, prioritize showcasing the value of certification, streamlining the purchase proces and offering incentives to active participants. By analyzing user behavior and addresing barriers, we can encourage more conversions from users already engaged in the program.

4.2 Box Plot for distribution of last completed step, step number and week

```
median_step
```

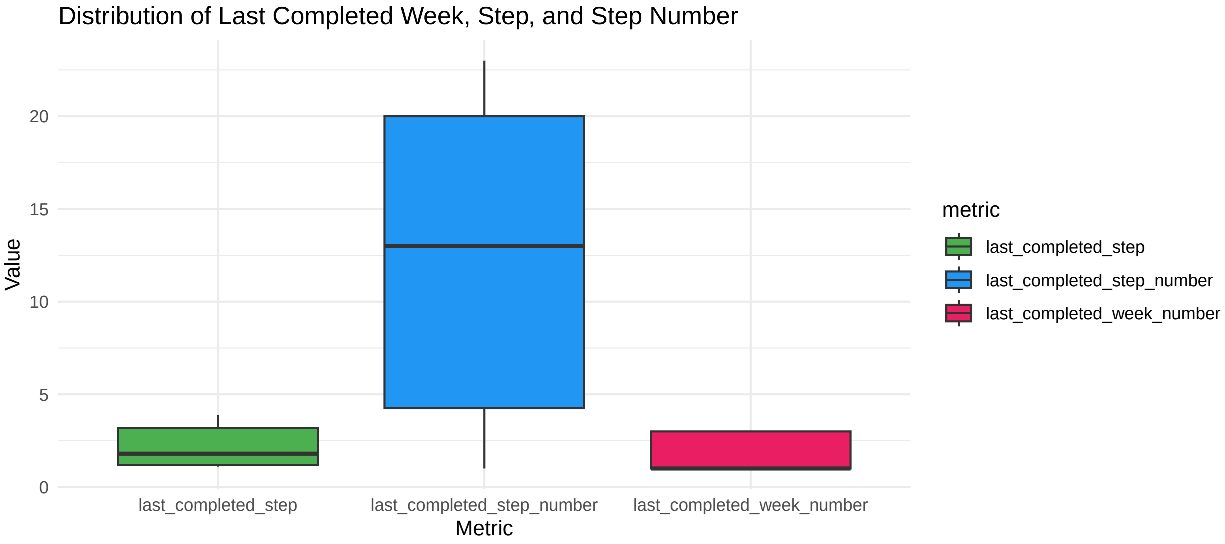
```
##   median_last_completed_step
## 1                        1.8
```

```
median_week
```

```
##   median_last_completed_week_number
## 1                                 1
```

```
median_step_number
```

```
##   median_last_completed_step_number
## 1                                13
```

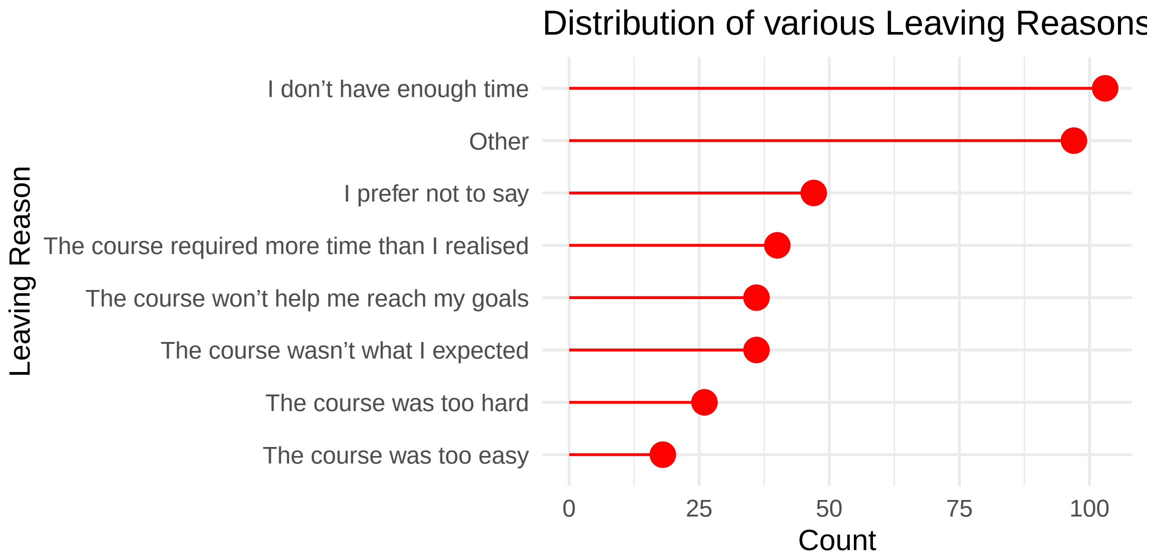


Median last completed step: 1.8 Median last completed week number: 1 Median last completed step number: 13

We see users are generally completing the program in the early stages, with the median user only reaching about the 13th step out of the full program.

This indicates potential opportunities to improve retention and encourage users to progress further through the content. This might also suggest the need of modifications to be done in the course to keep the learners equipped in order to achieve the objectives and goals.

4.3 Distribution of various leaving reasons with frequency



It appears that certain top reasons users leave the program, including "I don't have enough time" (103 users), "Other" unspecified reasons (97 users), and "I prefer not to say" (47 users).

This suggests opportunities to examine program structure, time commitments, and user expectations to enhance retention by addressing these common attrition drivers. This also suggests opportunities to examine program structure, time commitments, and user expectations to improve retention. Consider offering flexible scheduling options can be implemented. Provide extra support for users struggling with program difficulty (26 users) and analyze unspecified reasons like "Other" (97 users) to identify additional pain points. Regularly collect and incorporate user feedback to drive meaningful program improvements.

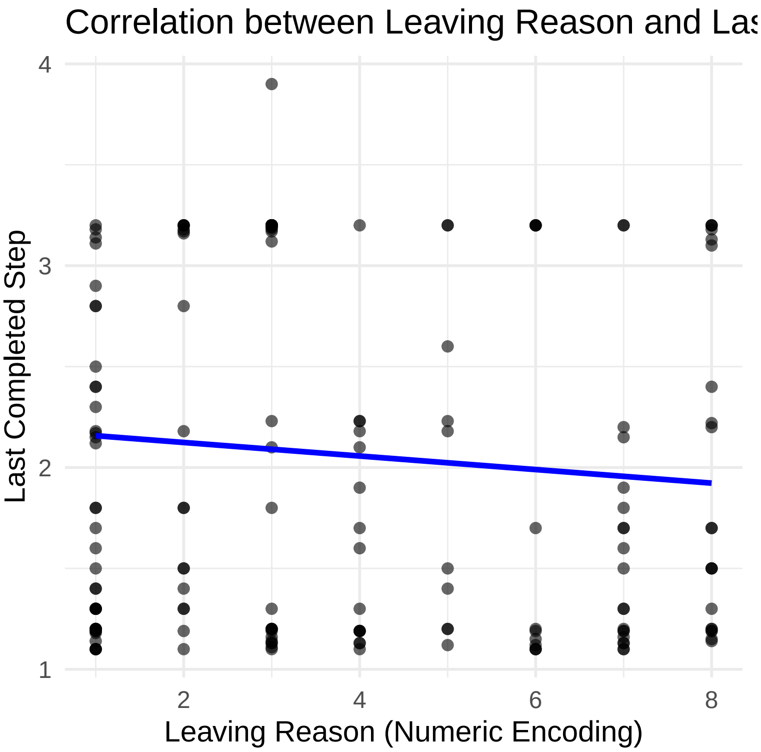
4.4 Correlation between Leaving responses

```
correlation_levresponse_lastcompstep
```

```
## [1] -0.09183262
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 217 rows containing non-finite outside the scale range
## (`stat_smooth()`).
## Warning: Removed 217 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



The data shows a moderate negative correlation of -0.09183262 between users' reasons for leaving the program and the last completed step. This suggests that users who progressed further in the program were less likely to report reasons for attrition, indicating engagement may be linked to lower dropout rates.

5. Evaluation

Overall, I can conclude that my exploratory data analysis on the dataset "Cyber Security: Safety At Home, Online, and in Life" highlights several key opportunities to improve program outcomes.

The 13.4% conversion rate for certificate purchases indicates room to optimize the user journey and increase conversions among engaged participants.Maybe prioritize showcasing the value of certification, streamlining the purchase proces and offering incentives to active participants. By analyzing user behavior and addresing barriers, we can encourage more conversions from users already engaged in the program.

The median's of last_step, step number and week number also say there is a potential need of opportunities to improve retention and encourage users to progress further through the content. This might also suggest the need of modifications to be done in the course to keep the learners equipped in order to achieve the objectives and goals.

Addressing common reasons for attrition, such as time constraints and unmet expectations, could boost retention.

The negative correlation between completed steps and reported leaving reasons suggests increased engagement may drive lower dropout rates. Incorporating user feedback to refine program structure, content, and support could enhance the course creator or the one who is running the course.

6. Deployment

For the deployment stage of CRISP-DM cycle I will be generating an analysis report and presentation to highlight my findings of the raw Futurelearn MOOC Dataset “Cyber Security: Safety At Home, Online, and in Life”