FutureLearn Analysis Report

Dimitrios Poulimenos - 200291237

Semester 1 - 2023/24

Introduction

This report aims to provide valuable insights into the online course titled "Cyber Security: Safety At Home, Online, and in Life," offered by Newcastle University and made accessible to the public through the online skills provider FutureLearn. In addition to analyzing the course's performance, this study delves into the pivotal role of learners' educational backgrounds in relation to their course completion rates. Furthermore, the report explores how this relationship has evolved over the course of seven runs.

The analysis presented in this report follows the Cross Industry Standard Process for Data Mining (CRISP-DM), a comprehensive framework that facilitates data-driven decision-making. To derive meaningful insights, this report conducts two cycles of the CRISP-DM model. These cycles enable us to produce in-depth findings and uncover trends that shed light on the interplay between educational backgrounds and course completion rates across multiple course runs.

Round 1 of the CRISP-DM Cycle

1. Business Understanding

To initiate the process effectively, it's essential to clearly outline the business objectives, desired outcomes, and establish concrete criteria for evaluating success. Once these foundational elements are in place, I can confidently advance to the subsequent stages while adhering to the structured Cross Industry Standard Process for Data Mining (CRISP-DM) framework.

1.1 Objective

The primary objective of this analysis is to gain valuable insights into the and dynamics of the "Cyber Security: Safety At Home, Online, and in Life" course, jointly provided by Newcastle University and FutureLearn. These insights aim to benefit educational institutions, course administrators, and online learning platforms by informing data-driven decisions.

The results are anticipated to provide actionable recommendations for course improvement, enhancing the overall learning experience. The insights are relevant not only to Newcastle University and FutureLearn but also to a broader audience of online course providers seeking to optimize their offerings.

1.2 Success Criteria

The success of this research hinges on several key criteria:

Data Quality: Utilize high-quality and accurate data sources to ensure the reliability and validity of the findings.

Objective Alignment: The analysis should directly contribute to the understanding and enhancement of the "Cyber Security: Safety At Home, Online, and in Life" course.

Accessibility: Present the results in a clear and accessible format to empower stakeholders to make informed decisions effortlessly.

Adhering to these success criteria, this analysis endeavors to offer tangible and practical insights for educational institutions and online learning platforms, ultimately enhancing the quality of online education.

Initial Research Question

After the objectives now we have to establish a question which this report aims to answer:

"How do the completion rates of learners with higher education levels compare to those with lower education levels across the seven course runs?"

1.3 Data Understanding

In this second phase of the first cycle, the objective is to provide a comprehensive overview of the data. This includes explaining the data's collection process, its composition, and the interrelationships between its various components. Additionally, this phase serves to define the subsequent steps to be taken in the project.

1.3.1 Data Colection

The data is sourced from the publicly accessible online course titled "Cyber Security: Safety At Home, Online, and in Life," which is offered by Newcastle University and hosted on the online skills platform, FutureLearn. This dataset is structured in CSV format and primarily encompasses participant-related information from all seven runs of the course.

1.3.2 Exploring the data

Following the data collection phase, a thorough analysis was conducted. This involved a comprehensive assessment of data quality, completeness, structure, and variable types. The dataset is comprised of eight distinct CSV files, each corresponding to a different course, and is accompanied by a PDF document outlining the course content.

It's essential to note that the composition of these files was not consistent across all cycles, leading to some variation in the data collection process. Consequently, for a seamless analysis, the decision was made to retain only those files that were present in all cycles. Specifically, the following two files were selected for analysis:

- **cyber-security-enrollments**: This file contains participant demographics and includes the following columns:
  - learner_id
  - enrolled_at
  - unenrolled_at
  - role
  - fully_participated_at
  - purchased_statement_at
  - gender
  - country

  - age_range
  - highest_education_level
  - employment_status
  - employment_area
  - detected_country

- **cyber-security-step-activity**: This file holds data concerning each participant's engagement with different course steps. It includes the following columns:

  - learner_id
  - step
  - week_number
  - step_number
  - first_visited_at
  - last_completed_at

By combining the information from these two selected files, it is anticipated that the desired results can be achieved for the analysis. The "cyber-security-enrollments" file provides essential participant demographic details, while the "cyber-security-step-activity" file offers insights into participant interactions with specific course steps. This strategic data selection ensures that the analysis is well-informed and aligns with the project's objectives.

1.4 Data Preparation

Continuing in the next phase of the first cycle we have the preparation of the data. In this phase, the cleaning, disaggregating, and organizing of the data takes place in such a way that we can create a dataset with which we can get the information we need for our analysis. It is important to note that this phase is the most crucial one because the whole analysis will be based on these data.

1.4.1 Data Cleansing

Initially, I began by eliminating rows with the "Unknown" value to ensure that I had complete and reliable data for subsequent analysis. After this, I proceeded to prune unnecessary columns from the enrolments datasets. For your reference, the columns that were removed from the "cyber-security-enrollments" files are as follows:

- unenrolled_at
- role
- fully_participated_at
- purchased_statement_at
- employment_status
- employment_area
- detected_country

Subsequently, I performed another data cleaning step by eliminating rows that lacked the "learner_id," which serves as the key for merging these two datasets In the "cyber-security-step-activity" files, I encountered a situation where multiple rows shared both the same "learner_id" and the same "step." This scenario was not expected, as in the course, each step should ideally be undertaken by a learner no more than once. Therefore, I took the necessary steps to address this issue and rectify the data accordingly. This step was crucial to ensure that the data remained well-structured and ready for further analysis.

1.4.2 Data Wrangling

After completing the data cleaning process, the next step involved data wrangling. Initially, I merged the seven "cyber-security-enrollments" files into a consolidated dataset. During this consolidation, I introduced a new column titled "Run," which was assigned values based on the originating file number. The same procedure was applied to the seven "cyber-security-step-activity" files, resulting in two final datasets, each prepared for merging.

The merging process was executed in a way that preserved the data that existed in both datasets, thereby ensuring that no rows contained non-existent variables. This careful merging strategy aimed to maintain data integrity and consistency across the resulting dataset.

Subsequently, I introduced a new column labeled "Completed" within the dataset. This column was designed to indicate whether a particular "learner_id" had successfully completed the corresponding step of the course. The concept behind this column was straightforward: if a learner had not completed any step of the entire course, it implied that they did not complete the course as a whole. Thus, the "Completed" column served as an informative indicator of course completion status for each "learner_id."With the given context, I added a new column called "Course_Status," which indicates whether the respective learner_id has successfully passed the course or not.

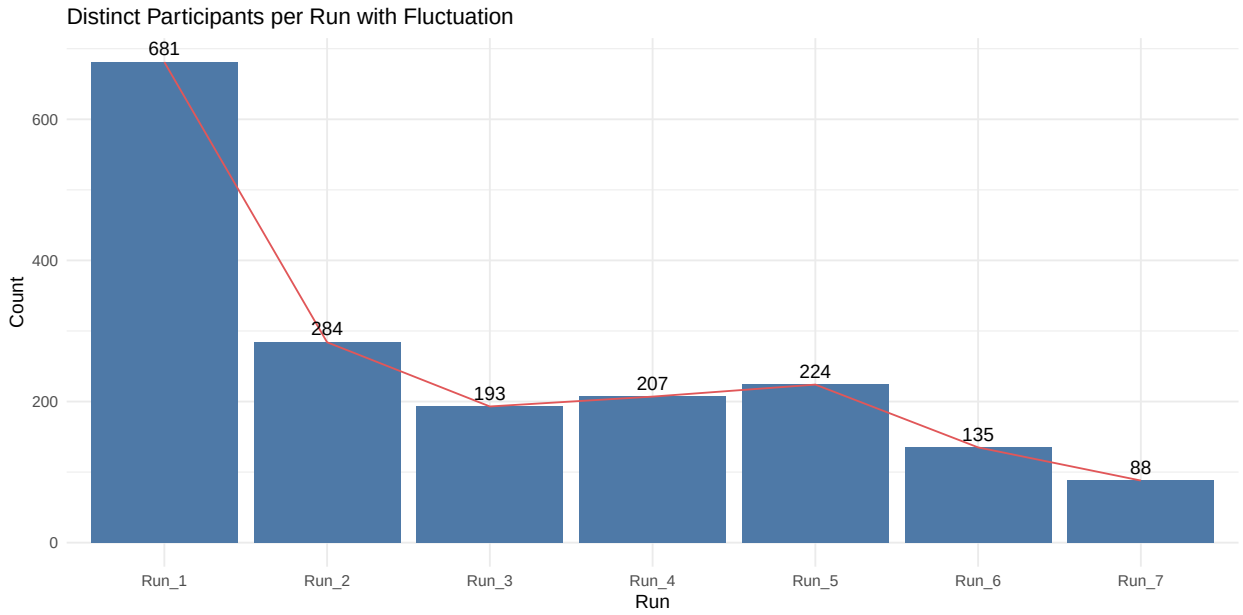
1.5 Modelling

With the data now thoroughly cleaned, formatted, and aligned with the initial objectives, the next phase involves conducting exploratory data analysis. This critical step entails visualizing the data to extract essential insights that will enable me to address my original research question effectively.

1.5.1 Exploratory Data Analysis (EDA)

The following analysis is designed to provide a visual representation of the data, enabling us to address our primary question and gain deeper insights into our dataset.

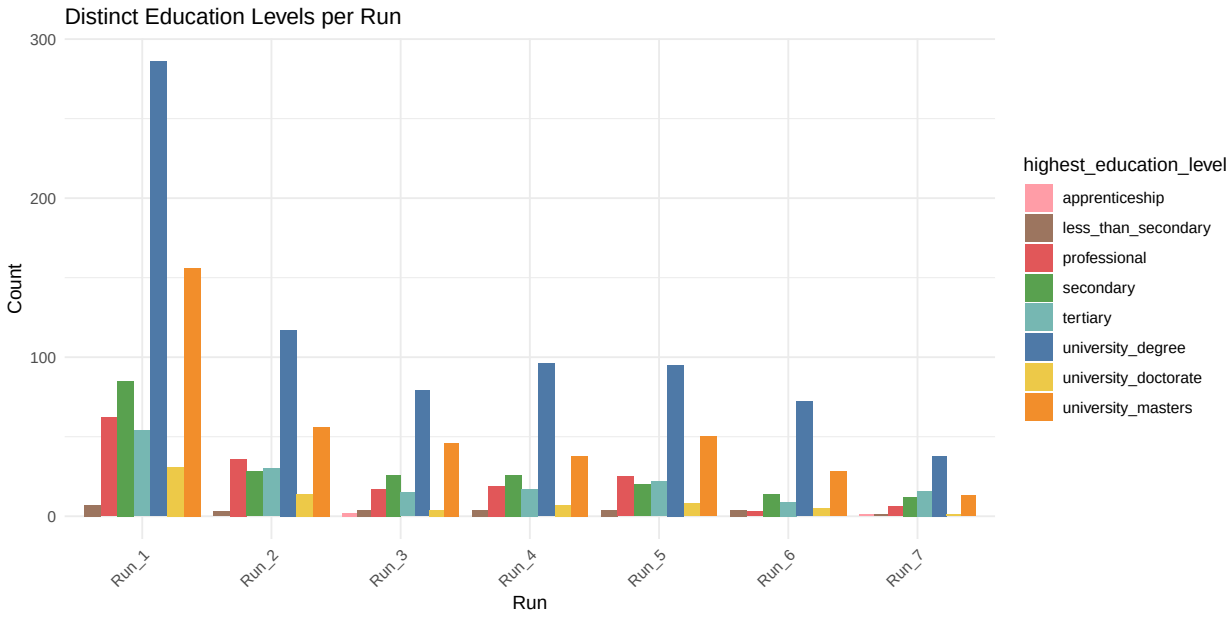
1.5.1.1 Distribution of participants per run



The above plot illustrates the participant distribution across the 7 runs of the course. Notably, the first run exhibits the highest level of participation among all 7 runs. As we progress through subsequent runs, there is a gradual decline in participation, marked by a brief increase in the 3rd run. However, starting from the 5th run, the participation decreases again, ultimately reaching its lowest point in the 7th run, which records the smallest number of participants.

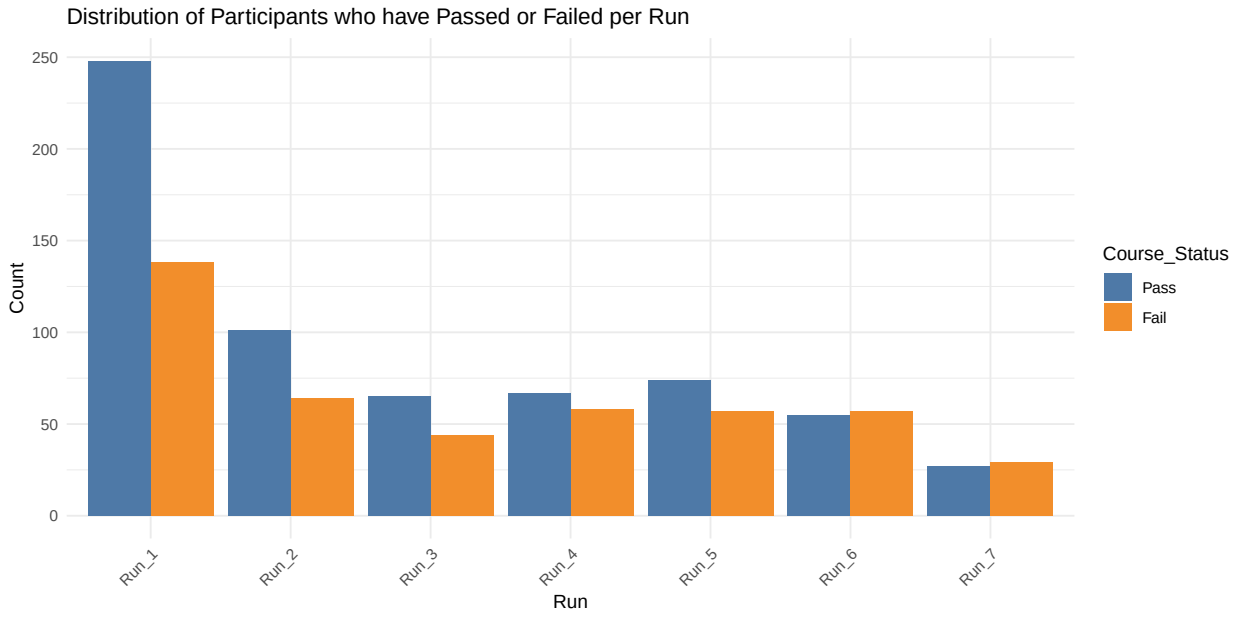
1.5.1.2 Distribution of the educational level per run

The following plot provides insight into the distribution of participants in each run, organized by their educational levels. Notably, individuals with a university degree consistently represent the highest percentage of participants across all runs. Following closely, participants with a master's degree typically occupy the second position, except for the final run where individuals with a tertiary degree take that spot. It's evident that these two educational levels play a predominant role in shaping overall participation trends, emphasizing their significance in the subsequent analyses.

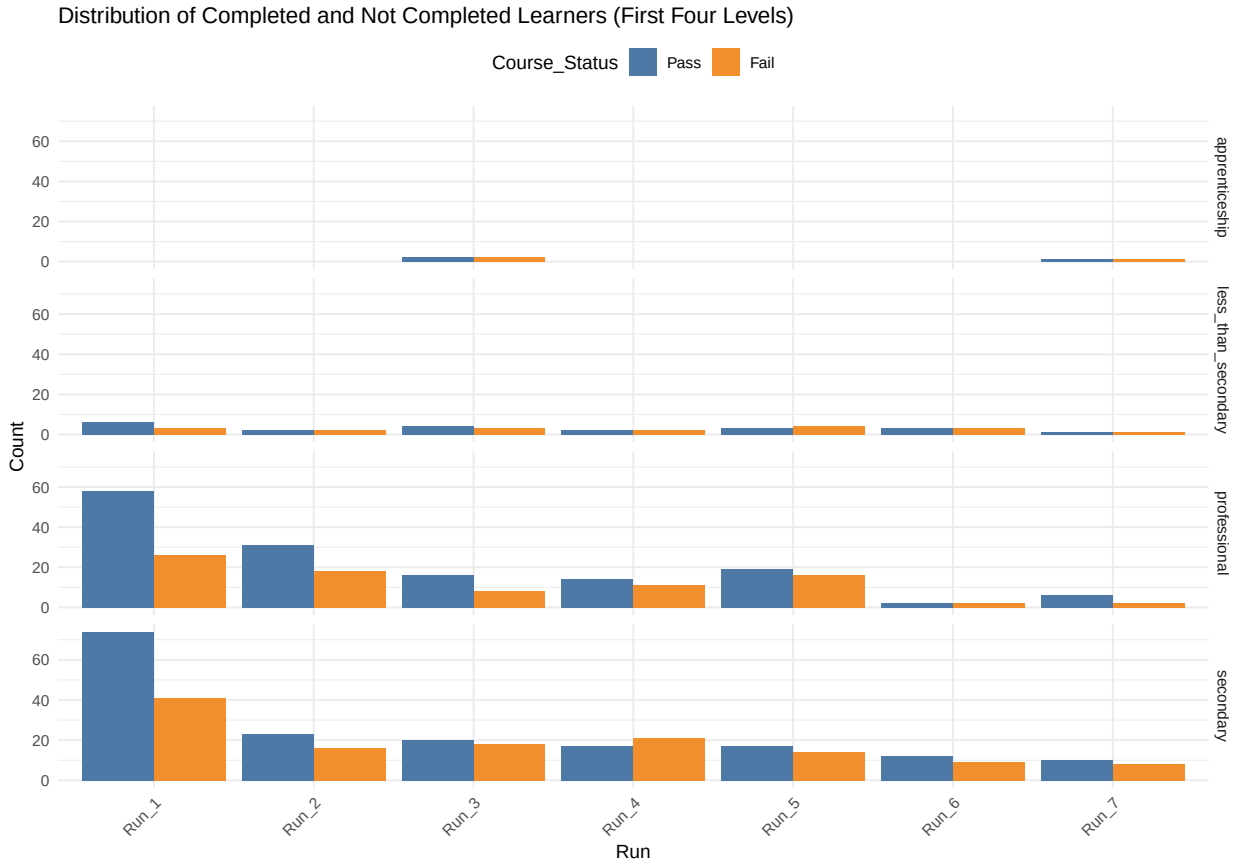


1.5.1.3 Distribution of the course completion per run

In the upcoming plot, we observe the participant distribution categorized by their course completion status in each run. Notably, in the first five runs of the course, the majority of participants successfully passed. However, in the final two runs, there's a reversal, where the percentage of participants who did not pass exceeds those who did. This observation suggests the possibility that the last two runs might have posed greater challenges compared to the preceding ones.



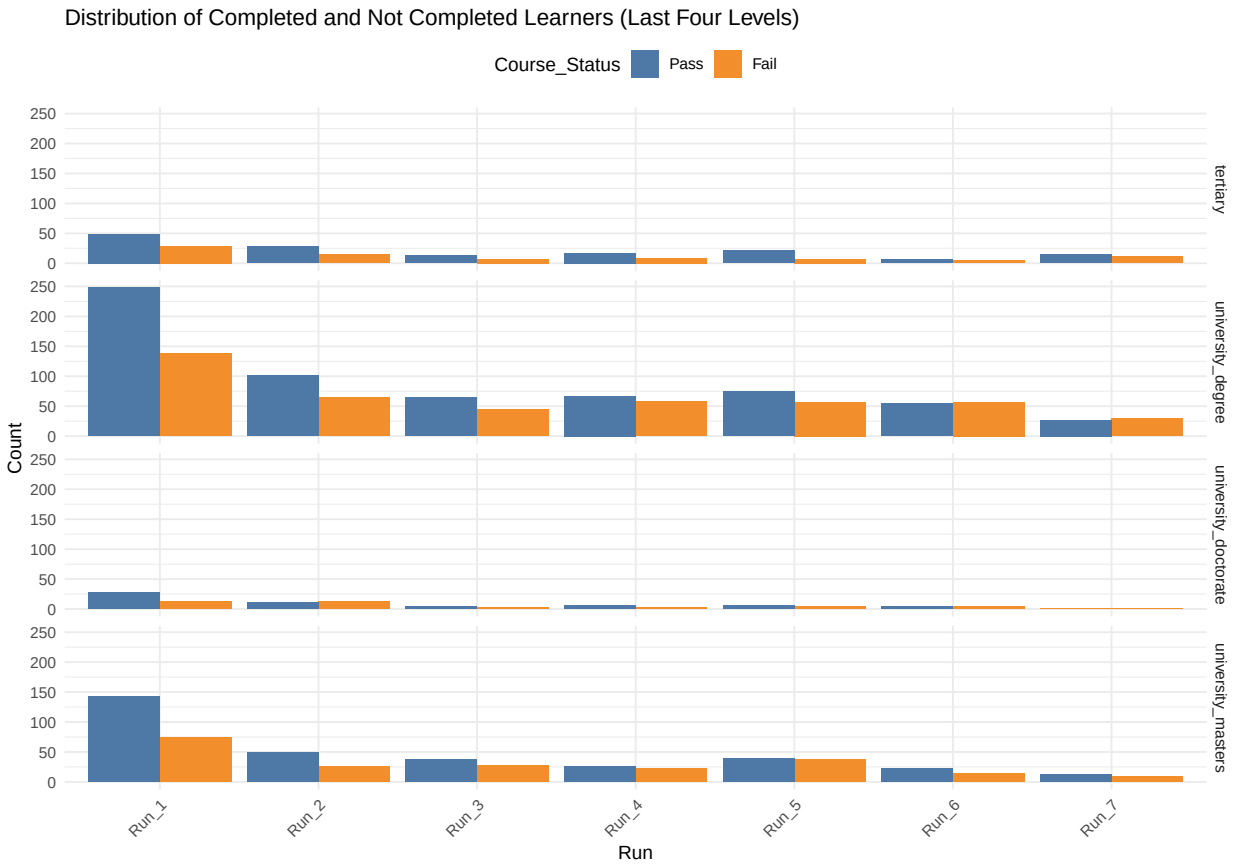
1.5.1.4 Comparing Pass and Fail of each education level per run



In the above plot, we delved into the distribution of participants from the first four education levels, uncov-

ering intriguing patterns and trends.

Participants with an apprenticeship background exclusively appear in runs 3 and 7. Notably, the success-failure ratio is evenly balanced at 50/50 in these runs, suggesting a diverse performance among apprenticeship-level participants. Moving to the "less_than_secondary" level, we encounter a varied landscape. In runs 1 and 3, the number of passers exceeds those who did not succeed. In runs 2, 4, and 6, the percentages of successful participants and non-passers are closely matched. However, run 6 deviates slightly with a higher percentage of non-passers. Participants at the "professional" level present a more consistent performance pattern. Run 6 stands out as the only run where the number of participants who passed the course is equal to those who did not. In contrast, in all other runs, the percentage of successful course completions is notably higher. Finally, participants at the "secondary" education level reveal a unique trend. Run 4 is the exception, where the percentage of non-passers surpasses those who successfully completed the course. In all other runs, participants who passed the course clearly outweigh those who did not.



Now, turning our attention to the last four education levels in the above plot, we uncover some intriguing observations: Participants with a tertiary education consistently display a positive trend. In all seven runs, the percentage of participants who successfully passed the course surpasses those who did not. Moving on to the "university_degree" level, we encounter some variations. In run 6, the percentage of participants who failed the course exceeds those who passed, while in run 7, the two groups are evenly balanced. However, in runs 1 to 5, the number of participants who successfully completed the course is greater than those who did not. Surprisingly, at the "university_doctorate" level, we observe distinct patterns. In run 7, an unusual scenario unfolds where everyone has failed the course. In runs 2, 3, and 6, the number of participants who passed and those who failed are nearly equal. In contrast, for runs 1, 4, and 5, the percentage of participants who passed the course significantly outweighs those who did not. Lastly, participants with

a "university_masters" level exhibit noteworthy trends. Only in run 5 do we find that the numbers of participants who passed and failed are nearly equal. In all other runs, the percentage of participants who successfully completed the course is notably higher.

These findings shed light on the performance of participants from all education levels across different runs, highlighting diverse outcomes and trends in course completion.

1.6 Evaluation

These results align with the objective and success criteria outlined at the beginning of the analysis. The primary objective was to gain insights into the dynamics of the "Cyber Security: Safety At Home, Online, and in Life" course and contribute valuable information for educational institutions, course administrators, and online learning platforms.

The success criteria established the need for high-quality data, alignment with the analysis objective, and clear communication of results. Specifically, The results directly address the main analysis question, providing insights into completion rates based on education levels across the seven course runs. Moreover, the success criteria emphasized utilizing high-quality and accurate data sources. While the details of the data quality checks aren't explicitly mentioned in the response, it is implied that the analysis is based on reliable data, as the insights are presented with confidence. Also, the results are communicated in clear and accessible language, meeting the success criterion of presenting findings in a way that empowers stakeholders to make informed decisions effortlessly. Furthermore, the analysis highlights potential challenges in the last two runs, contributing actionable insights that may lead to improvements in course design and delivery. This aligns with the success criteria of providing practical recommendations for course improvement. And finally, the mention of potential challenges in the last two runs suggests a forward-looking perspective, aligning with the success criterion that emphasizes recommendations for future work, such as exploring specific patterns and incorporating feedback loops for continuous improvement.

Concluding, the analysis shows varying patterns. People with higher education, like university or master's degrees, generally had more course completions. However, there were some changes in certain runs, and in the last two runs, more people did not finish the course, suggesting it might have been tougher. Overall, this analysis helps me understand how completion rates differ among learners with different education levels across the course runs.

Round 2 of the CRISP-DM Cycle

2. Business Understanding

In the ongoing second cycle of CRISP-DM, the overarching objectives remain unchanged. Leveraging insights obtained in the first cycle, particularly regarding the success and failure rates in each course run categorized by educational levels, provides a foundation for extending our analysis. Building upon the previous findings, the subsequent question emerges:

"In the last two runs, characterized by the lowest completion rates, were there more male or female participants who successfully completed the 'Cyber Security: Safety At Home, Online, and in Life' course?"

2.1 Data Understanding

After a thorough examination of the existing data from the previous cycle, I have determined that additional data beyond what I already have will not be necessary. The selected data sources, particularly the two files from each run containing participant demographic information and their course completion status, are

deemed ideal for addressing the specific question in this cycle. The comprehensive nature of these pre-existing datasets ensures that I have all the required information to conduct the upcoming cycle of analysis.

2.2 Data Preparation

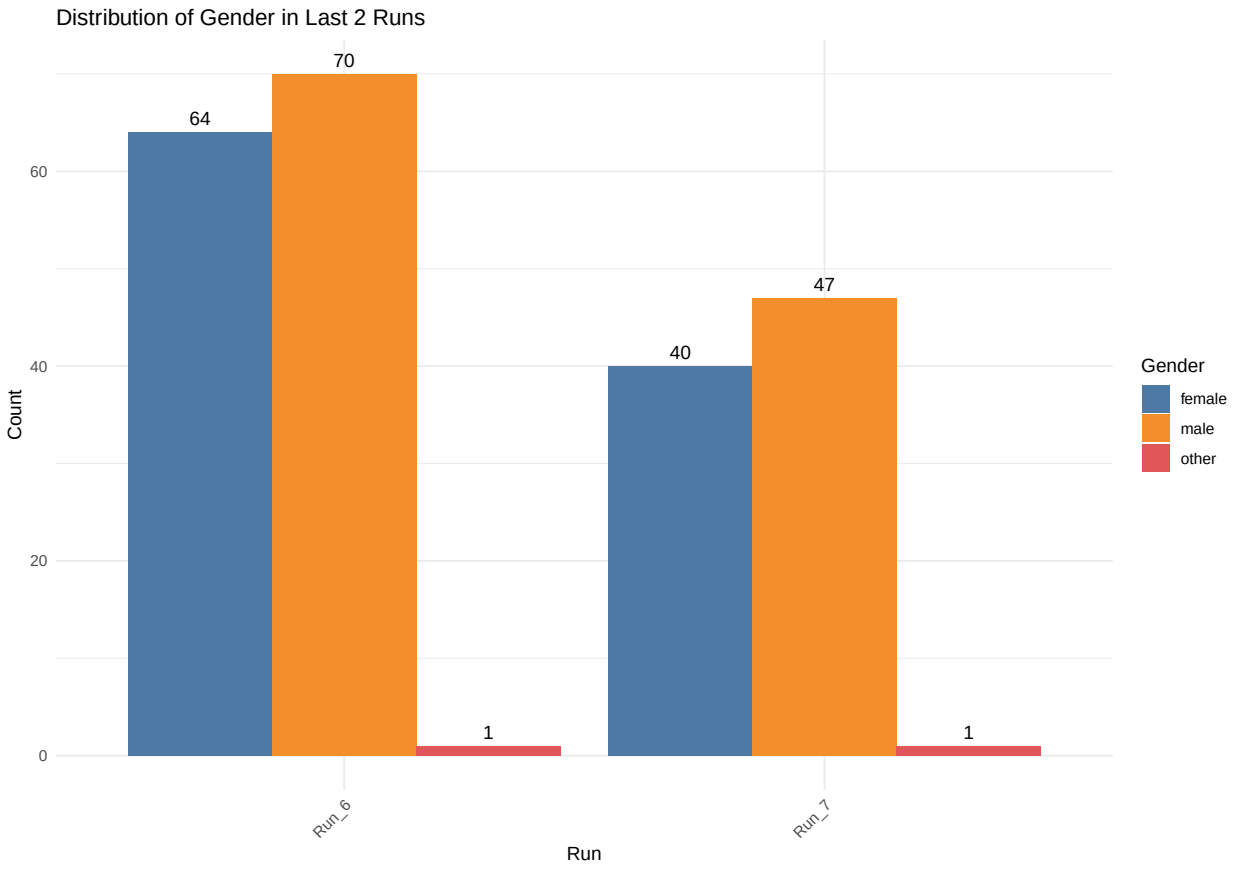
In this second cycle phase, the required data for analysis has been prepped and refined through cleaning and parameterization. In the initial cycle, a strategic decision was made to retain the "gender" column, anticipating its relevance for the second question in this cycle. However, I had to create a new column based on the last created column named Course_Status which have the name "Overall_Status" in order to make sure that I am getting the distinct number of learners for each gender. This deliberate choice not only aligns with the current analytical needs but also streamlines the data for subsequent phases. By retaining the necessary variables and proactively addressing specific future questions, the data processing workflow has been optimized, allowing for more focused and efficient analysis.

2.3 Modelling

2.3.1 Exploratory Data Analysis (EDA)

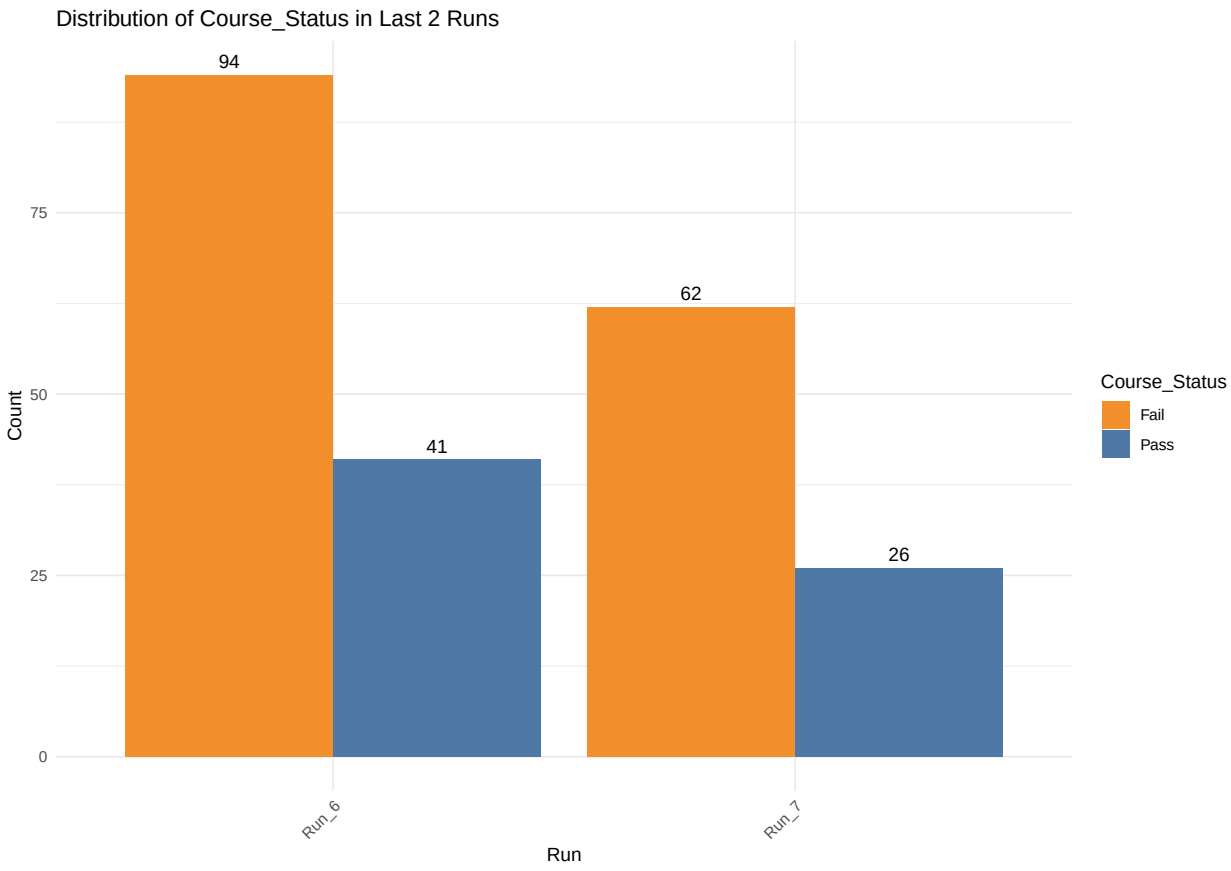
Embarking on the new exploratory data analysis, my focus is on visualizing the gender distribution in correlation with the success and failure rates during the last two runs of the course.

2.3.1.1 Gender distribution in the last two course runs



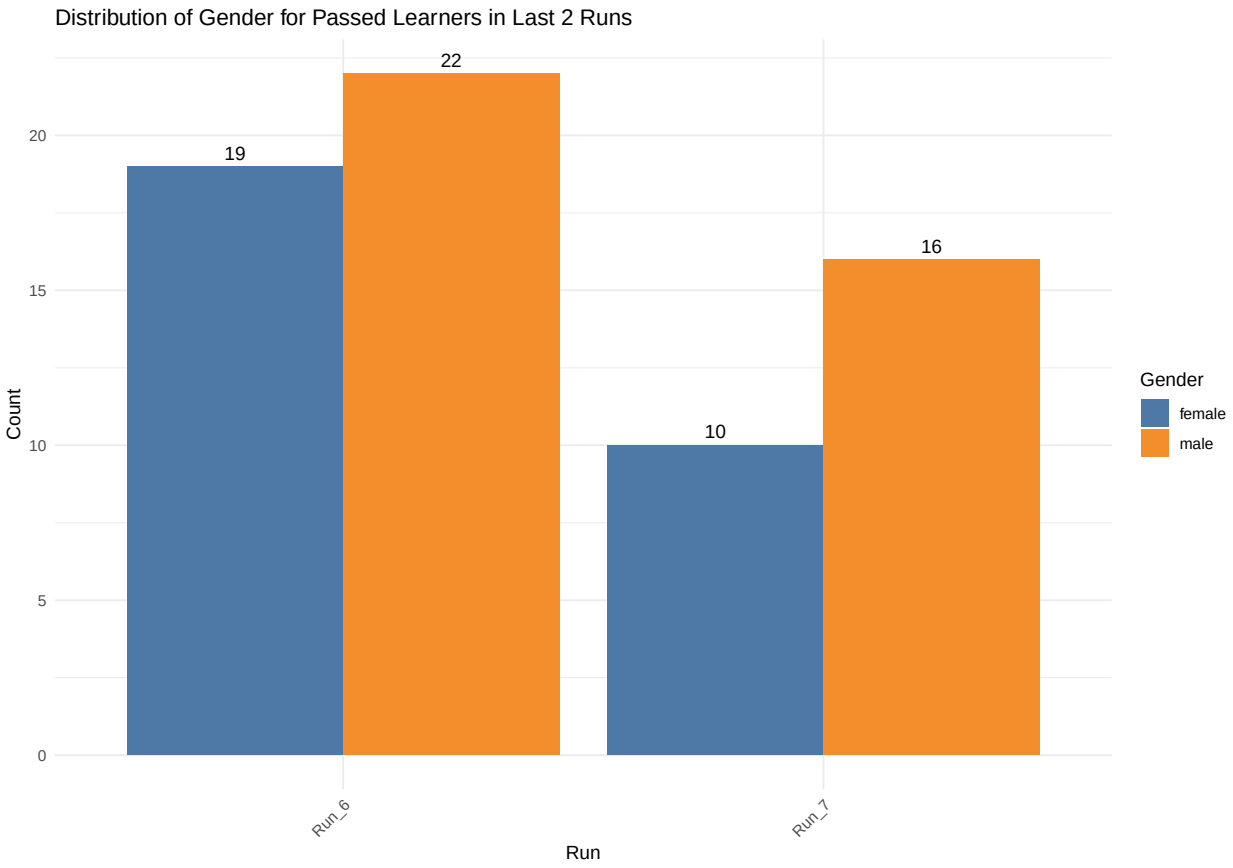
Commencing with a broad overview of participant distribution by gender, it becomes apparent that the male participants outnumber their female counterparts in both instances. Notably, there is a solitary registration that has designated "other" as the gender category; however, due to its minimal presence with just one registration in each case, we consider it inconsequential for our analytical purposes and thus will not factor it into our analysis.

2.3.1.2 Distribution of Pass/Fail in the last two course runs



In the above plot, a clear pattern emerges, affirming our earlier findings from the initial cycle. Notably, the number of failures significantly outweighs the number of successes in both the 6th and 7th course runs. This robust confirmation underscores the success of our initial analysis, providing a compelling foundation for further exploration of our second research question.

2.3.1.3 Pass/Fail Distribution according to gender in the last two course runs



After visualizing the distribution of successful participants based on gender in the last two runs, a discernible trend emerges, the count of male participants slightly surpasses that of female participants. This observation effectively addresses the pivotal question of the second cycle of the CRISP-DM methodology.

2.4 Evaluation

Concluding of both CRISP-DM cycles, the gleaned insights offer invaluable perspectives on the "Cyber Security: Safety At Home, Online, and in Life" course. Overall, I am confident in asserting that the initial objectives and success criteria have been successfully achieved. Furthermore, the outcomes of the analysis emerge as a valuable resource for stakeholders, equipping them to discern specific patterns and trends within the data, thereby enhancing their understanding of the course dynamics.

3. Deployment

Concluding the final phase encompassing both cycles of the CRISP-DM methodology, which involves the deployment of my findings to stakeholders, it is crucial to highlight that this report, coupled with an accompanying presentation, has been thoughtfully developed. The core aim is to inform stakeholders about the detailed insights gleaned from the analysis. Furthermore, the internal documentation is designed to serve as a valuable reference for future course configurations, underscoring the strategic importance of the gathered findings.