Learning Analytics | FutureLearn MOOC Dataset

Vaishnavi Jeurkar
Student No.: 230595217
Newcastle University, Newcastle upon Tyne

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. The insights derived from this analysis hold the potential to inform the course's learning design and facilitate intervention processes for at-risk students. These 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

The first phase of the cycle is to identify key stakeholders, define clear business objectives and success criteria.

1.1 Background

The integration of Learning Analytics into the cybersecurity MOOC by Newcastle University is a strategic initiative aimed at optimizing the learning environment. Learning Analytics, a key facet of Data Science, involves the systematic measurement, analysis, and reporting of learner data to enhance educational outcomes.

Traditional methods, like attendance monitoring, often fall short in capturing nuanced online engagement. By employing the CRISP-DM methodology, this initiative unifies diverse data sources—ranging from on-campus facility usage to VLE interactions. The goal is to provide instructors and course designers with actionable insights for informed decision-making, intervention strategies, and ongoing course refinement. This approach reflects a commitment to data-driven innovation and continual improvement in online education.

1.2 Business Objective 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. By understanding and optimizing the learning experience and the course environment, the insights derived from the data will inform enhancements and improvements to the educational platform.

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 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

"Can we identify the patterns of student engagement in the 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.

1.6 Constraints The primary constraint for this deliverable is time, as the analysis is conducted within a tight 2-week timeframe, limiting the in-depth exploration of the business intricacies.

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 patterns and trends in student learning, enabling the identification of successful teaching strategies and potential areas for improvement. 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 Following tools showcase a robust and well rounded approach to data analysis,

1. GitHub:
   Strengths: GitHub 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 commentary 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 profiles of every learner and steps of how they progressed through the course.

2.2 Data Files Description Following Data files can be viewed in the course runs.
1. Archetype Survey Response: This file 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 stamps.
3. Leaving Survey Response: This datafile captures the leaving reasons and step activity of learners who unenrolled from the course.
4. Question Response: This file 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 datafile 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 file 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 file captures comments and ratings of different learners for each week of the course.

2.3 Data Exploration We want to understand how students engage with the course. Whether it's reading articles, watching videos, or taking quizzes, students move through the course in steps. We're trying to find out if there are any patterns in how students complete these steps. In my initial look at the data, there are some files that have information about students and instructors, 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 Weekly Sentiment Survey file is only there for the 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 datafiles 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 enrolled or unenrolled in the course which would have been helpful to identify the dropout rate of the course but after inspecting further it is seen that it has many null entries for the unenrolled time stamp. For this reason I will not be using this file for analysis. The incomplete data of Archetype survey response, Weekly sentiment survey response and Leaving survey response makes it difficult to draw reliable insights from them. The only files with substantial amount of data are Step activity, Video stats and question response data. Step activity data provides a comprehensive overview of how learners progress through the steps and hence I will be using this file to observe the student engagement patterns in the course.

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 step activity file. This particular file provides a detailed overview of the timelines for each learner, indicating when they initiated and completed specific steps throughout the course.
Following features are extracted from the file:

```
## Rows: 28,304
## Columns: 7
## $ learner_id        <chr> "77454a73-6b8b-46a2-8dee-35f36b6c4fc1", "20e6ec35-0f~
## $ step              <dbl> 1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1.~
## $ week_number       <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ step_number       <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ first_visited_at  <chr> "2018-08-10 08:39:26 UTC", "2018-09-05 13:57:38 UTC"~
## $ last_completed_at <chr> NA, NA, "2018-09-10 00:53:16 UTC", NA, NA, "2018-09-~
## $ time_to_complete  <dbl> NA, NA, 48.6666667, NA, NA, 5.4333333, 9.2500000, NA~
```

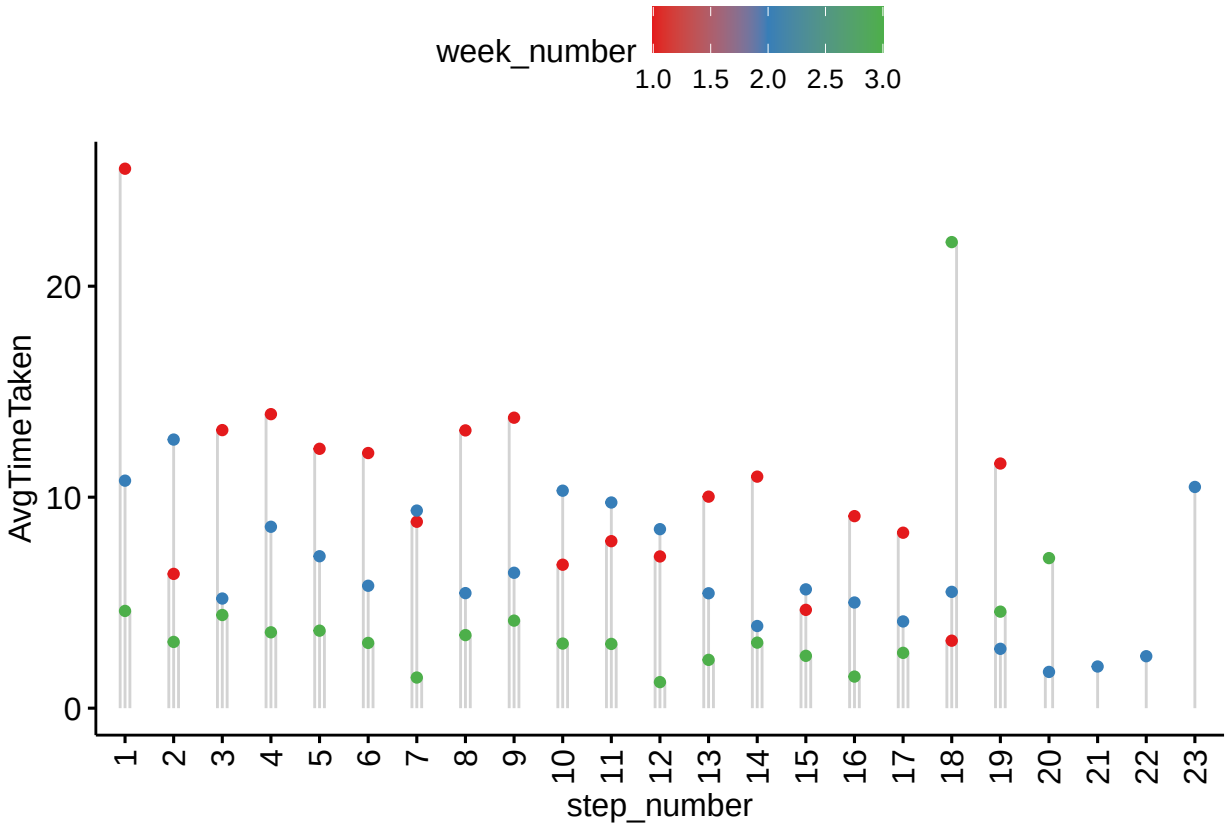
Given the presence of null values in the "last completed at" feature, we operate under the assumption that these null values signify instances where learners did not complete the corresponding step. Leveraging this assumption, we can identify the steps where the majority of learners have left the course incomplete.
Upon delving 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. The data collected in the 7th run of the course is well-structured with no missing or blank files. Given the recentness of this data, the majority of my analysis will be focused on the 7th run of the course.

3.2 Data Construction and Formatting The dataset contains step activity data for 1626 learners, with multiple recorded steps for each learner ID. The only column with null values is 'last_completed_at,' which is crucial for determining incomplete steps. As these null values are needed for analyzing incomplete steps, no data cleaning is required. There are a total of 3204 null string values in the 'last_completed_at' column, which have been replaced by 'NA' for consistency. By calculating the difference between timestamps, the 'time_to_complete' column is derived, representing the time taken to complete each step in minutes. 'Avg-TimeTaken' denotes the average time taken by all learners to complete a specific step. 'PercentIncomplete' indicates the percentage of learners who left a step incomplete. The 'result' variable captures the value of the last completed step number for each learner_id.

4. Modelling.

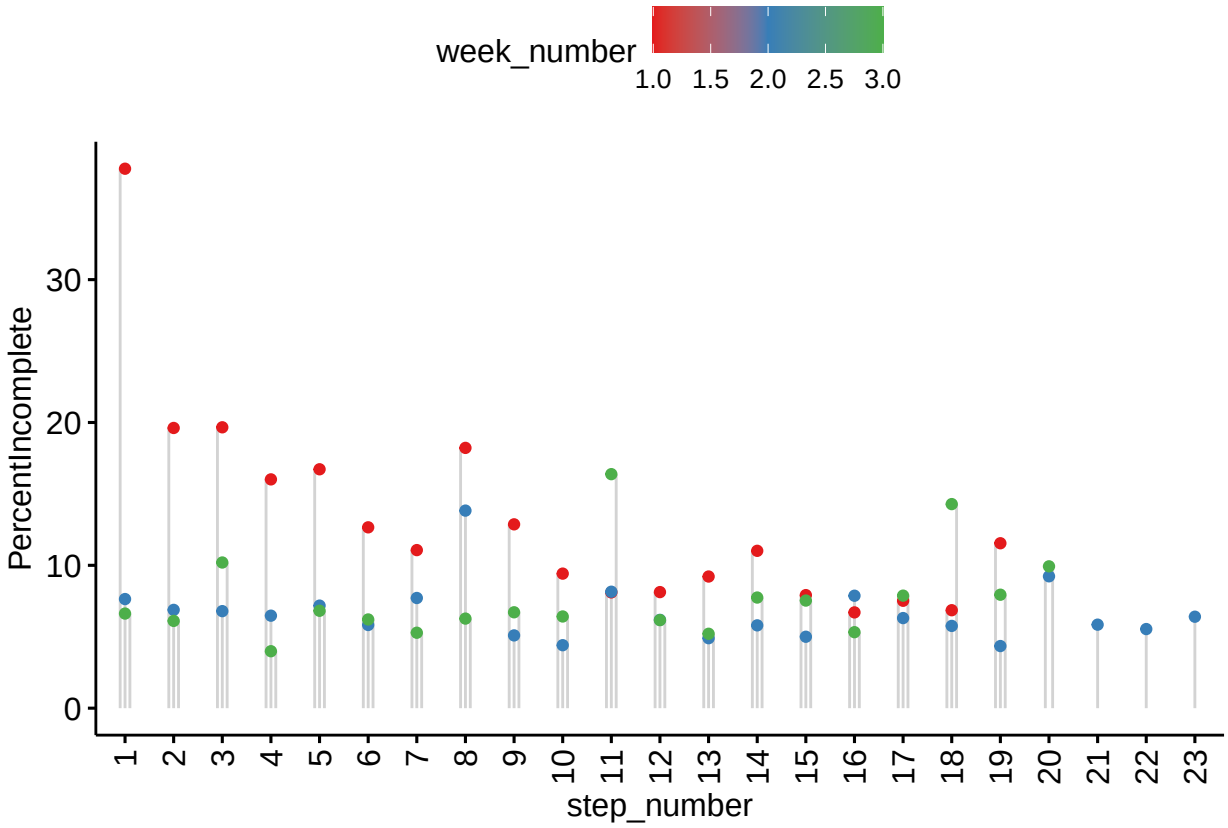
Upon concluding the data preparation phase, I revisited the earlier stages to verify adherence to the initial plan and assess potential impacts on the upcoming analysis. Finding no deviations or issues, I proceeded to complete the exploratory data analysis to answer our question - "Can we identify the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

4.1 Average time taken by student to complete the steps in each week.



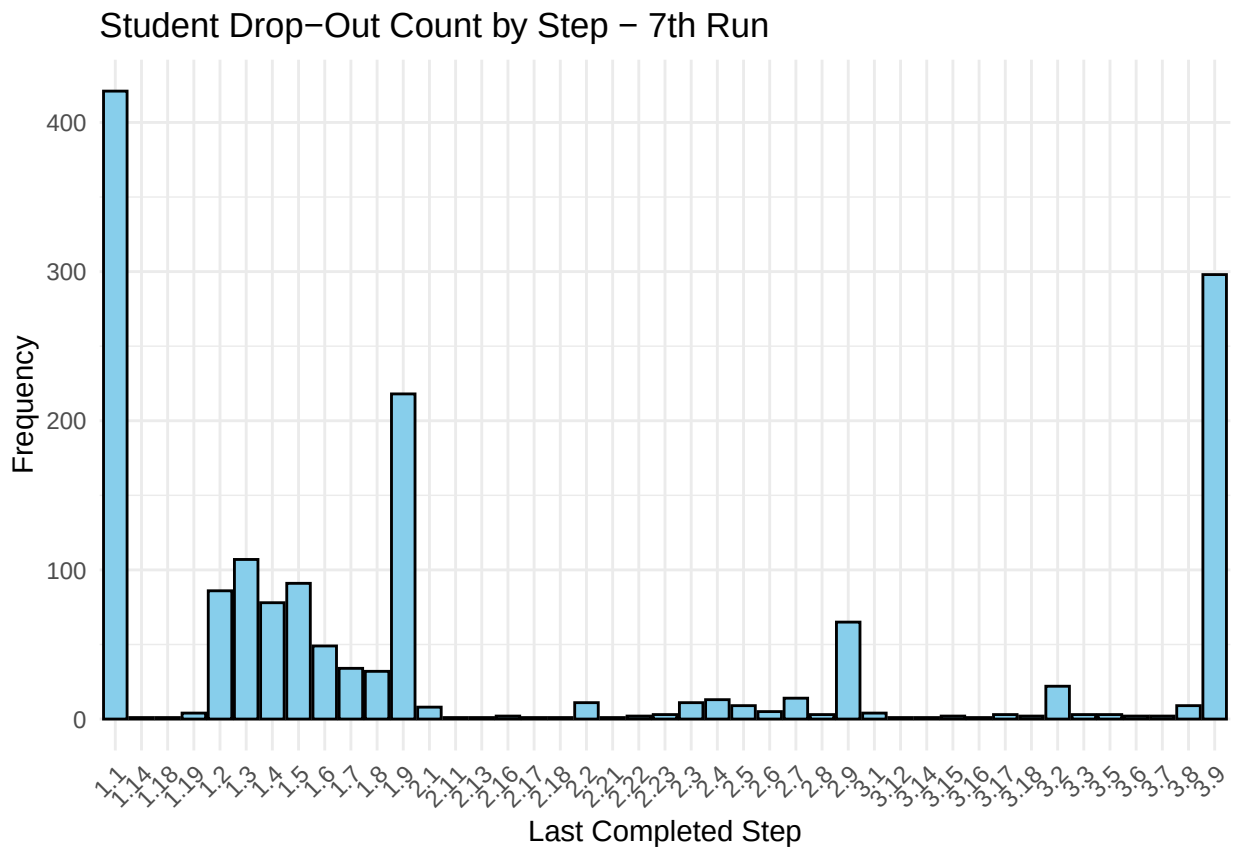
The analysis reveals interesting patterns in student behavior. Notably, students invest significant time in completing the initial step, the welcome video. Subsequent steps in the first week also exhibit longer completion times, possibly indicating a lower level of focus at the beginning of the course. As students progress, the steps in Week 2 show a decrease in completion time, reflecting growing interest and consistency in the learning process. Week 3 steps, on the other hand, demonstrate the shortest completion times, suggesting a heightened level of engagement. It's worth noting a spike in average time for step 18 in Week 3, possibly attributed to the increased difficulty of this particular quiz.

4.2 Percentage of students who left the step incomplete.



The visual representation illustrates a notable trend wherein a significant proportion of students leave the first-week steps incomplete. Approximately 37% of students do not engage with the welcome video, and over 15% do not complete steps 2 to 5 in Week 1. Notably, steps involving quizzes, such as 1.8, 2.8, 2.20, 3.11, and 3.18, exhibit an increased percentage of incompletion. This pattern suggests that students may encounter challenges in solving these quizzes or have a lower level of interest in these particular activities.

4.3 Calculating the student drop-out rate

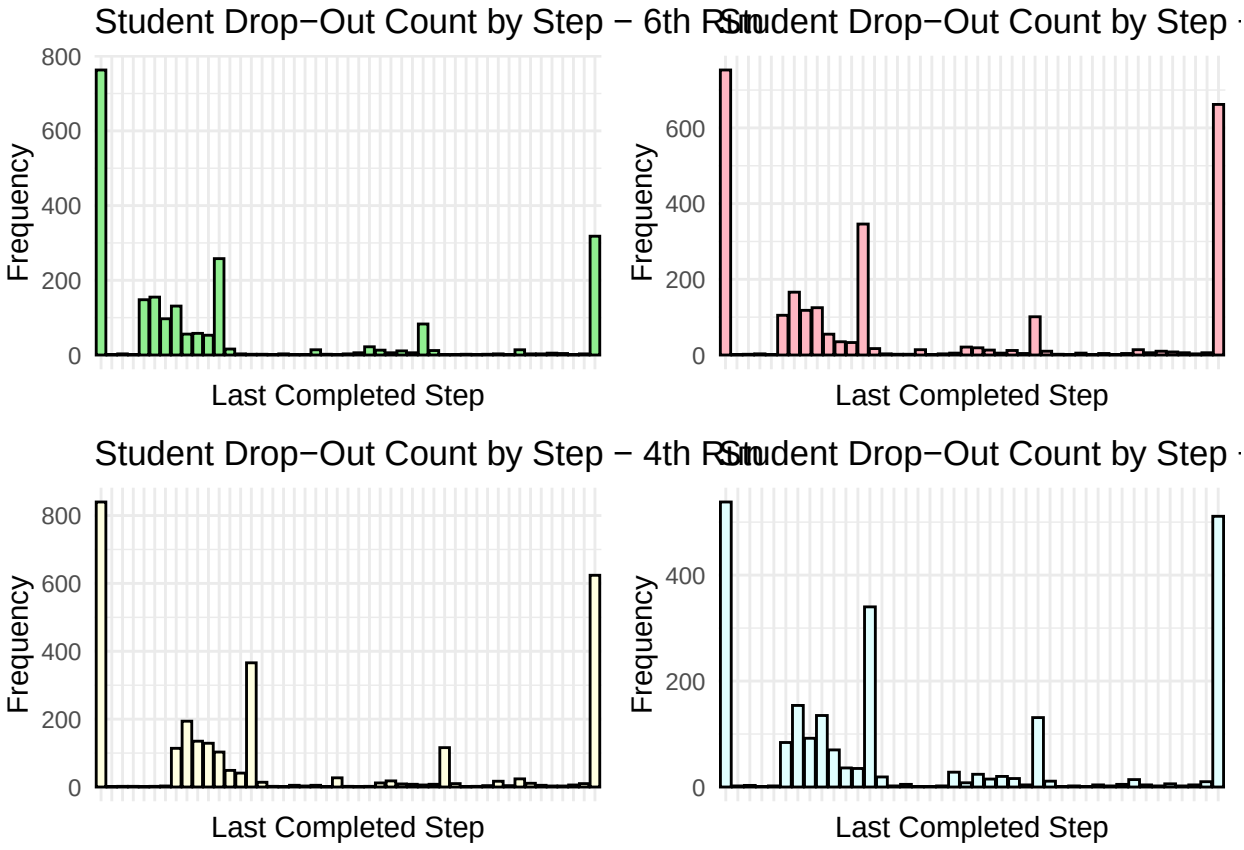


The analysis of the last completed step for each learner reveals interesting insights into the course engagement and attrition patterns. Notably, over 400 students exit the course immediately after the introduction video, indicating potential disinterest or a lack of engagement in the initial content. The dropout rates peak during the early steps, particularly from 1.1 to 1.5, suggesting a need for enhancing the appeal and engagement of these introductory sections.

A significant dropout point occurs at step 1.9, following the completion of the quiz at step 1.8. This could be attributed to the quiz's difficulty level, emphasizing the importance of assessing and adjusting the challenge posed by assessments. A similar trend is observed at step 2.9, aligning with the challenging quiz at step 2.8.

The overall completion to the last step (3.20) is relatively low, with approximately 300 students discontinuing the course by step 3.9. This insight underscores the importance of continuous improvement strategies to enhance course content, quizzes, and overall engagement throughout the learning journey.

4.4 Checking for replicability for drop-out count rate



The depicted plot illustrates the frequency of the last completed step for four different course runs (Run 3, 4, 5, and 6). A notable observation is the consistency in dropout patterns across all four runs, despite variations in the enrolled student numbers. This uniformity in dropout trends suggests that the issue is not solely influenced by the course size but rather indicates a need for a comprehensive revision of course content.

5. Evaluation

In this conclusive phase, the analysis of the FutureLearn cybersecurity course data has successfully met the defined business objectives and success criteria. It has delivered valuable insights into student engagement patterns, dropout tendencies, and areas for potential improvement. By exploring last completed steps, average time taken, and completion percentages based on week and step, 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 sufficient to capture general trends. Looking ahead, the plan is to conduct an in-depth analysis of student engagement and performance with different types of course material, aiming to gain further insights into the factors influencing student dropout rates and varying levels of engagement.

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 engagement patterns by scrutinizing diverse learning

materials in the cybersecurity course. Our primary questions for this phase include investigating the correlation between student interactions with course videos and their progression through steps. Additionally, we aim to explore how question performance contributes to overall engagement and identify patterns in specific content types (videos, questions) that significantly impact student dropout rates or sustained engagement. 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 into student engagement patterns with different learning materials. Success criteria include identifying clear correlations between video statistics, question performance, and step activity. 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 Costs

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. Analytical Tools: Utilizing sophisticated 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 materials and student engagement can inform strategic course development for future offerings.

2. Data Understanding

After defining clear business objectives I moved onto the phase 2 of second cycle.

2.1 Data Description In this analysis, two key data files from the 7th run of the course were utilized: the "video_stats" file and the "question_response" file. The "video_stats" file contains insightful information such as the step position of videos, video duration, total views, downloads, the percentage of the video viewed by learners, and views across different devices and continents. This data provides a comprehensive overview of student interactions with the course videos. On the other hand, the "question_response" file contributes to understanding student performance patterns in relation to their engagement. Key features in this file include quiz question number, step number, week number, question response, and learner id. Analyzing this data will aid in identifying reasons for student dropout after specific quizzes, offering valuable insights into the dynamics of student engagement and learning outcomes.

```
## [1] "Features of video stats file"

## Rows: 13
## Columns: 28
## $ step_position                  <chr> "1.1", "1.14", "1.17", "1.19", "1.5", "~
## $ title                          <chr> "Welcome to the course", "Why would any~
## $ video_duration                 <int> 99, 362, 241, 348, 281, 37, 312, 92, 42~
## $ total_views                    <int> 1041, 489, 362, 476, 777, 345, 282, 270~
## $ total_downloads                <int> 43, 16, 21, 21, 55, 6, 15, 11, 15, 7, 1~
## $ total_caption_views            <int> 14, 8, 8, 4, 12, 5, 7, 4, 5, 2, 3, 3, 2
## $ total_transcript_views         <int> 196, 112, 75, 102, 164, 67, 58, 54, 73,~
## $ viewed_hd                      <int> 41, 15, 11, 10, 20, 3, 6, 5, 8, 4, 3, 2~
## $ Five                           <dbl> 80.88, 73.62, 81.77, 69.75, 76.45, 78.2~
## $ Ten                            <dbl> 79.63, 71.78, 79.28, 66.81, 72.59, 77.3~
## $ TwentyFive                     <dbl> 76.18, 68.92, 74.31, 61.76, 65.51, 75.9~
## $ Fifty                          <dbl> 72.43, 64.62, 69.61, 57.56, 58.94, 74.7~
## $ SeventyFive                    <dbl> 69.84, 61.35, 67.13, 55.46, 55.60, 73.6~
## $ NinetyFive                     <dbl> 68.30, 60.33, 62.71, 53.57, 53.15, 71.8~
## $ Hundred                        <dbl> 66.28, 57.46, 49.72, 46.85, 44.92, 71.0~
## $ console_device_percentage      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
## $ desktop_device_percentage      <dbl> 75.31, 82.00, 79.83, 80.04, 79.79, 80.0~
## $ mobile_device_percentage       <dbl> 20.46, 10.63, 10.77, 11.13, 13.90, 10.7~
## $ tv_device_percentage           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
## $ tablet_device_percentage       <dbl> 4.03, 7.16, 9.12, 8.19, 6.05, 8.70, 10.~
## $ unknown_device_percentage      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
## $ europe_views_percentage        <dbl> 52.16, 63.39, 62.15, 64.08, 59.59, 62.9~
## $ oceania_views_percentage       <dbl> 2.79, 4.70, 4.70, 3.57, 4.12, 4.35, 4.6~
## $ asia_views_percentage          <dbl> 25.55, 17.18, 17.13, 16.39, 20.21, 17.6~
## $ north_america_views_percentage <dbl> 8.07, 7.57, 7.73, 7.56, 6.69, 6.96, 10.~
## $ south_america_views_percentage <dbl> 2.31, 2.04, 2.49, 2.94, 2.96, 2.61, 2.4~
## $ africa_views_percentage        <dbl> 8.65, 4.50, 5.25, 4.62, 5.79, 4.35, 7.0~
## $ antarctica_views_percentage    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0

## [1] "Features of question response file"

## Rows: 9,927
## Columns: 10
## $ learner_id      <chr> "77454a73-6b8b-46a2-8dee-35f36b6c4fc1", "62449cd5-916b~
## $ quiz_question   <chr> "1.8.1", "1.8.1", "1.8.1", "1.8.1", "1.8.1", "1.8.1", ~
## $ question_type   <chr> "MultipleChoice", "MultipleChoice", "MultipleChoice", ~
## $ week_number     <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ step_number     <int> 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, ~
## $ question_number <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ response        <chr> "1,2,3", "1,2", "1,2", "1,2,3", "3", "1,2", "1,2,3", "~
## $ cloze_response  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ submitted_at    <chr> "2018-07-31 15:44:17 UTC", "2018-09-10 02:16:21 UTC", ~
## $ correct         <chr> "true", "false", "false", "true", "false", "false", "t~
```

2.2 Data Quality To ensure the integrity of the analysis, it was necessary to address null values in the learner ID column of the quiz question file. Rows with blank learner IDs 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. The step_position feature was transformed into a character variable. Specific columns related to viewer percentages were renamed to ensure the proper visibility of x-axis labels. In the question response file, rows with null values in the learner ID column were removed. These formatting adjustments contribute to the overall readability and effectiveness of the subsequent visualizations.

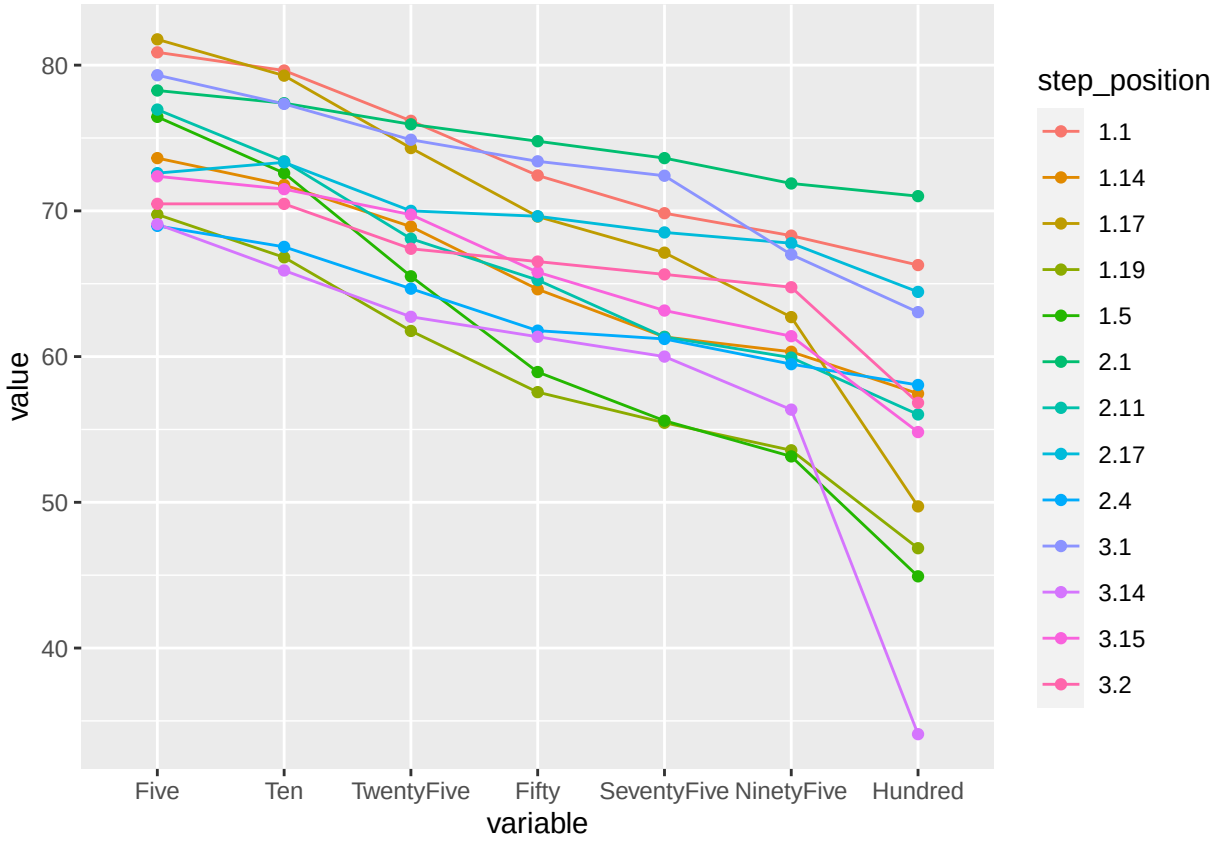
4 Modelling

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

4.1 Check the correlation between video duration and total views The aim was to assess the potential relationship between video duration and viewer engagement rate, accomplished by calculating the Pearson's correlation coefficient for these variables. The computed value of -0.03 revealed a weak negative correlation. This signifies that while there's a slight tendency for the viewer's engagement rate to slightly decrease with longer video durations, this association is notably faint. The Pearson's correlation coefficient assesses the strength and direction of a linear relationship between two variables, with this result indicating a negligible negative correlation between video duration and viewer engagement rate.

```
## [1] "Correlation value between total views and video duration is -0.03326797174516"
```

4.2 To find patterns in video viewership of learner's at various steps



The parallel coordinates plot illustrates the variation in video viewership percentage across step positions, spanning from 5 percent to 100 percent viewed by learners. Notably, the video at step 3.14, titled "Exploring security: biometric authentication," shows a sharp decline and garners the least complete viewership. It's noticeable that shorter videos, such as those at step positions 2.1 and 1.1, are more frequently viewed entirely, though a declining trend is observed across these steps.

4.3 To find the correlation between step completion and student performance in quizes

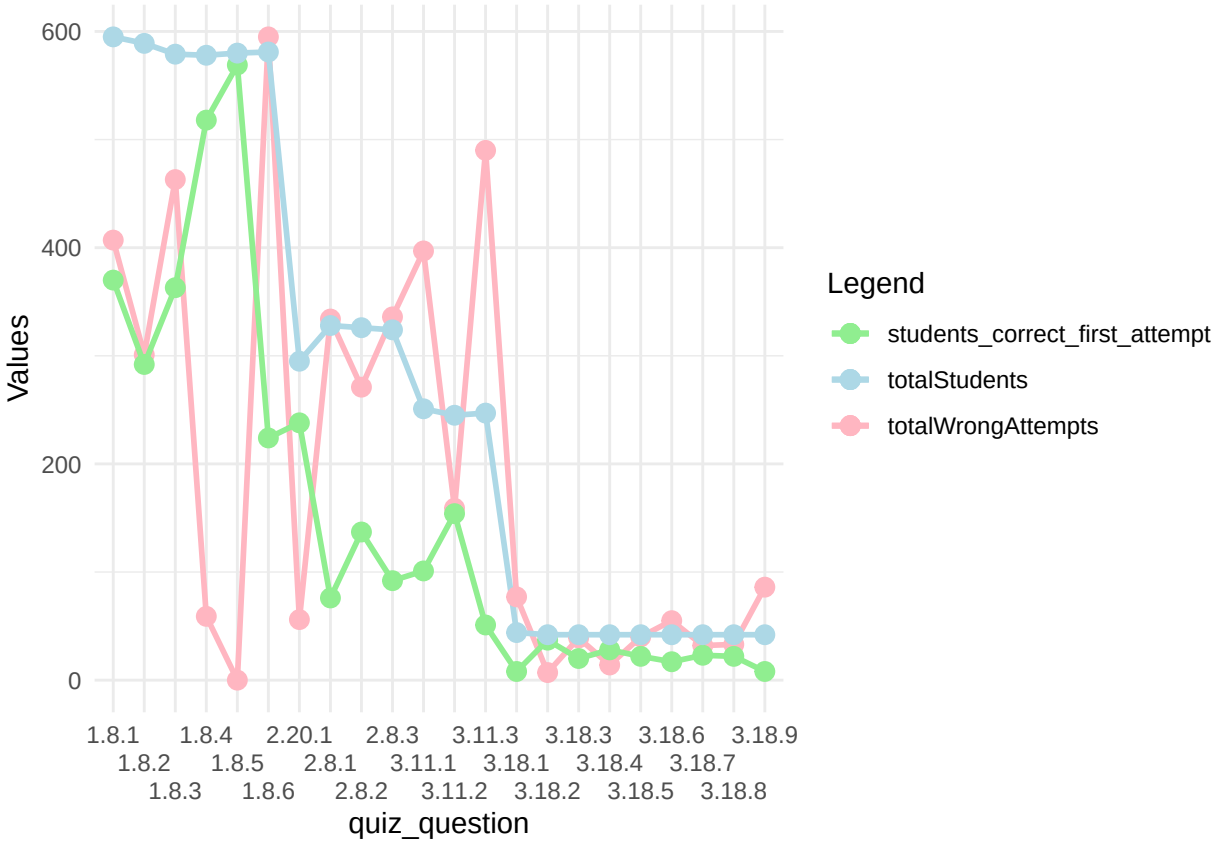
```
##                 week step_number correct_complete incorrect_complete
## Week 1 Step 8      1           8                3                16
## Week 2 Step 8      2           8                0                 0
## Week 2 Step 20     2          20                0                 0
## Week 3 Step 11     3          11                0                 2
## Week 3 Step 18     3          18                0                 1
```

To evaluate student engagement and pinpoint potential reasons for dropout following quiz completion, a detailed analysis of quiz performance was imperative. Leveraging data from both the step_activity_data and question_response datasets, the presented table outlines the week and step numbers corresponding to the quizzes, accompanied by key metrics for correct and incorrect completions. The 'correct_complete' column reflects the count of learners who not only finished the steps preceding the quiz but also answered all questions correctly on their initial attempt. Conversely, 'incorrect_complete' showcases the number of students who completed all steps but provided incorrect answers to all quiz questions. For example, in quiz 1.8, only 3 students achieved a perfect score on their first attempt, while notably, 16 students completed the course content yet answered the quiz incorrectly. The relatively low number of students successfully completing all steps and answering all questions correctly raises concerns, especially with none achieving

this in the week 2 and 3 quizzes. This discrepancy could indicate dissatisfaction or misunderstandings among students, potentially contributing to dropout occurrences at steps 1.8 and 2.8.

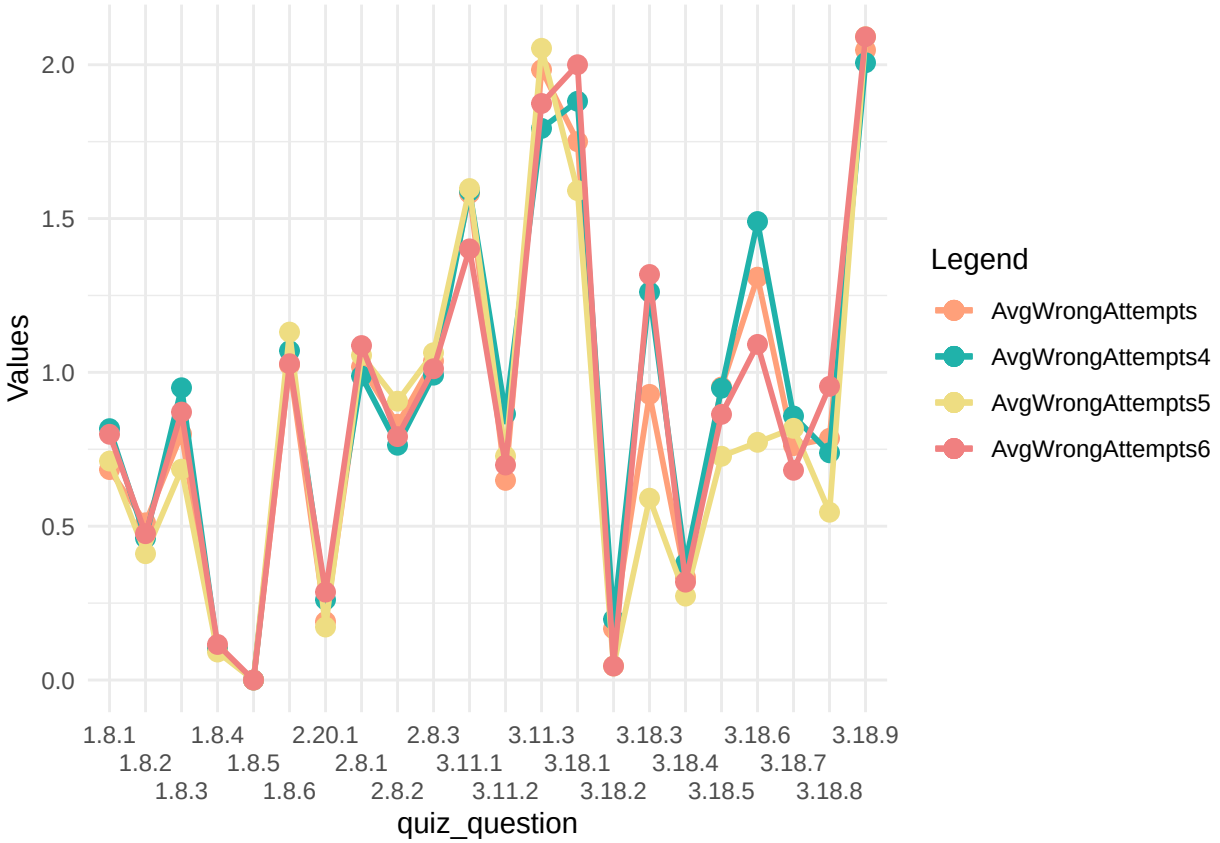
4.4 Finding patterns in quiz completion steps.

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



The line chart provides a comprehensive comparison across three key metrics: ‘totalStudents’ - signifying the total number of quiz attempts, ‘totalWrongAttempts’ - tallying the overall incorrect attempts by each learner, and ‘students_correct_first_attempt’ - indicating learners who answered correctly on their first try. A noticeable trend emerges from week 1 to week 3, with a sharp decline in quiz attempts, plummeting from 600 students in week 1 to merely 50 in week 3. Among the questions, 1.8.6 received the maximum attempts, while question 1.8.5 saw nearly all students providing correct answers on their first attempt. However, challenges surfaced in week 3, notably in questions 3.11.1 and 3.11.3, where fewer students achieved correct responses on their initial try. Intriguingly, the questions for quiz 3.18 seem more approachable, considering the closer alignment between the count of students who answered correctly on their first attempt and the total number of attempts.

4.5 Checking difficulty level across questions for different runs.



I've conducted an analysis to assess the variance in question difficulty for students across multiple runs (from run 4 to 7). This involved collating data on the total wrong answers attributed to each learner ID within these runs. Subsequently, I computed the average number of wrong answers across all learner IDs, aiming to gauge the average level of difficulty encountered by students across these runs.

It appears that certain questions, specifically 3.11.3 and 3.18.3, exhibit variability in difficulty levels across different runs. Notably, learners in run 4 and 6 encountered challenges in question 3.18.1, while those in run 5 and 7 faced difficulties in question 3.11.3. Students from run 5 find 3.18.6 to be less challenging than students from other runs. However, aside from these variations, most questions exhibit consistent trends across the four runs. Questions like 1.8.4, 1.8.5, and 3.18.2 seem relatively easier, whereas question 3.18.9 appears to be the most challenging. Additionally, there's an observable pattern of an increasing trend in wrong answers provided by students from week one to week three.

5. Evaluation

In its essence, the second round does an in depth analysis on video engagement and question performance which In summary, the second phase of analysis delved deeply into video engagement and question performance, shedding light on crucial aspects impacting student success and course attrition. Initial observations revealed a weak negative correlation between video length and viewership, suggesting the need for a closer examination of video content and viewer preferences. The parallel coordinates plot highlighted viewer engagement patterns, offering valuable insights into video consumption trends among learners, aiding instructors in understanding student interests and preferences.

Analyzing question responses from the 7th course run exposed challenges regarding student success and dropout tendencies post-quiz completion. Few students successfully answered all questions on their first attempt after completing the course content, indicating potential issues with content relevance or quiz

difficulty. Notably, some students answered quiz questions incorrectly despite finishing the course content, potentially indicating dissatisfaction and contributing to dropout rates.

Further investigation into question difficulty unveiled declining student participation, hinting at disinterest or early attrition in the course. However, limitations in the available data, notably decreasing learner counts throughout quizzes, hindered comprehensive insights into question difficulty. Analysis across multiple course runs revealed consistent question difficulty levels, suggesting a need for course refinement to enhance teaching methodologies, content relevance, and student engagement in quizzes. Addressing these findings is crucial to improve student performance, reduce dropout rates at challenging steps, and enhance viewer engagement with course videos.

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"

References

1. Futzing and Moseying: Interviews with Professional Data Analysts on Exploration Practices, IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 25, NO. 1, JANUARY 2019, Sara Alspaugh and Nava Zokaei and Andrea Liu and Cindy Jin and Marti A. Hearst.
2. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research Nicole Gray Weiskopf, Chunhua Weng.
3. https://www.ibm.com/docs/en/spss-modeler/18.2.0?topic=guide-introduction-crisp-dm
4. R Markdown: The Definitive Guide, Yihui Xie, J. J. Allaire, Garrett Grolemund