FutureLearn Course Analysis

Zheng Luo

17 November, 2023

Introduction

As the landscape of online education continues to evolve rapidly, the need for data-driven insights into course effectiveness and learner engagement becomes increasingly paramount. This project, centered around the analysis of the MOOC "Cyber Security: Safety At Home, Online, and in Life," aims to harness the power of data science to unravel the intricacies of learner behavior, course appeal, and overall effectiveness. Utilizing the structured approach of the CRISP-DM framework, this analysis delves into multiple aspects of the course, ranging from enrollment trends and learner demographics to detailed feedback on course content and structure. The primary goal is to distill actionable insights that can guide course improvements, enhance learner engagement, and ultimately contribute to the success of the online learning experience. By meticulously examining both quantitative and qualitative data across two comprehensive cycles of analysis, this project endeavors to provide a multi-dimensional view of the course's performance, offering strategic recommendations for its evolution in the dynamic world of online education.

To improve the clarity and impact of my data visualizations in this report, I have:

- Ensured each table and graph has clear titles and labels for better understanding.
- Adopted color schemes accessible to readers with color vision deficiencies.
- Included concise interpretations under each visualization to guide readers through the insights.

Round 1 of the CRISP-DM Cycle

Business Understanding

The onset of this analysis process involves defining clear objectives for business use, identifying the main question, and setting out success criteria against which the results will be evaluated. Based on the criteria established in this initial phase, a plan will be developed within the framework of CRISP-DM. This structured approach ensures that the analysis is not only methodical but also tailored to meet the specific needs and expectations of the learners.

Objective

For the analysis to be considered successful, it must utilize high-quality data to yield credible insights. The findings should directly address the primary objective and answer the core research question, while being presented in a manner that is accessible and easily interpretable by readers.

This cycle aimed to unravel the nuances of the course dynamics, including aspects like enrollment consistency, the diversity of the learner bas, and learner performance in quizzes, Understanding these elements is crucial for assessing the course's overall impact and identifying potential strategies for enhancement.

Guided by the overarching goal, the central question that this report seeks to address is:

"How has the course quality developed over the past 7 runs"

Data Understanding

In the first phase of the analysis, the focus is the key metrics that could shed light on the course's effectiveness. This also included the quality of the data: its requirements, availability, and source reliability are key considerations to meet the objectives and generate meaningful insights. This phase is critical to ensure that the chosen data aligns with the objectives and that the goals are realistic given the available data.

Collect Initial Data

For this project, data was sourced from an online educational platform that provided detailed records of learner engagement for the "Cyber Security: Safety At Home, Online, and in Life" MOOC. The data encompasses various metrics that reflect student participation and progress through the course. The comprehensive dataset allowed for a granular analysis of learner engagement over seven course iterations.

Explore Data

An in-depth exploration of the dataset was conducted next, where potential data quality issues were identified, variables and their types were examined, and the alignment of data availability with the needs of the study was evaluated. This preliminary review helped in mitigating the risk of investing time in a dataset that might not serve the analysis requirements, ensuring a systematic approach to validate its usefulness.

Initial findings revealed that while some data were consistently recorded across all course iterations, such as "enrollment", "question responses" and "step activities". The consistent data points provided a relatively solid foundation for analysis.

To summarize, the raw data encompassed the following key metrics across various years of the course offerings:

- Enrollment data (collected from 7 runs)
- Question response data (collected from 7 runs)
- Step activity data (collected from 7 runs)

Despite the variations in data collection over the years, the available information was deemed sufficient for conducting a robust analysis. The reliability of the recorded data was established, with the understanding that some data might require careful handling during the preparation phase due to inconsistencies in formats.

Concluding this phase, it was determined that the data set held valuable insights that could be mined to meet the analysis objectives and success criteria without needing to alter the initial scope. Hence, the decision was made to proceed to the more granular stages of data preparation and analysis.

Data Preparation

The subsequent phase involved preparing the data for analysis, which included cleansing, transforming, and selecting the relevant data points. The goal is to normalize the dataset to ensure consistency and accuracy, thereby facilitating a smoother transition to the modeling phase. This preparation was essential for enhancing the reliability of the forthcoming analysis and for ensuring that the results would be comprehensible and verifiable by others.

Integrate Data

The datasets from the seven course runs were integrated using a tailored function "merge_run_data", using the common column "learner_id" of all the files mentioned earlier. This ensured that all learner-related information was cohesively combined into one data-set, facilitating a unified analysis.

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
id.x
responded_at
archetype
id.y
left_at
leaving_reason
last_completed_step_at
last_completed_step
last_completed_week_number
last_completed_step_number
quiz_question
question_type
week_number
step_number
question_number
response
cloze_response
submitted_at
correct
run_number

Construct Data

Generated records included summarisation of enrollment trends and gender distribution, providing a macro view of the course's reach and demographic balance.

Derived attributes such as the proportion of correct quiz responses ("proportion_correct") and the time taken to complete specific steps ("time_diff") were generated to evaluate learner performance and engagement.

Clean Data

The data cleaning process involved standardizing the "learner_id" field across all datasets to ensure consistency. The function "check_and_convert_learner_id" was employed to convert "learner_id" to character strings where necessary.

Additionally, the quiz responses, originally recorded as 'true' or 'false', were standardized to integer values (1 for correct, 0 for incorrect) to facilitate quantitative analysis of learner performance.

Format Data

Date and time fields were reformatted using the "lubridate" package to facilitate temporal analysis of step activity, such as calculating the average time taken to complete steps and determining step completion percentages.

Data-set

The final data-set, "combined_data", represents a comprehensive aggregation of learner interactions across seven iterations of the course. It includes enrollment details, step activity, and quiz responses. This data-set was further processed to produce "average_time_data", which provided insights into the time taken to complete course steps and the completion rates. I picked the course steps "1.1", "2.1", "3.1" in all 7 runs to compare the changes of average time learners spent.

Table 2: Small summary table of average time learners spent in selected course step(using first 2 runs as an example)

run_number	1	1	1	2	2	2
step	1.1	2.1	3.1	1.1	2.1	3.1
average_time	45.522974	28.372761	23.090169	15.588742	7.970595	5.536811
completion_percentage	82.32126	94.83181	94.57627	79.34647	94.48068	94.81627

This data-set is the foundation for subsequent analysis, designed to uncover insights into learner behavior and course effectiveness over time.

Modelling

In this phase of the analysis, I employed visual data modeling techniques with the help of the package "ggplot2", mainly focusing on line plots, bar charts, and percentage charts. These methods were chosen for their effectiveness in revealing trends and patterns in enrollment, gender distribution, quiz performance, and step completion rate and times.

Enrollment Trend Analysis:

A line plot with blue lines and red points, displaying enrollment numbers across seven course runs.

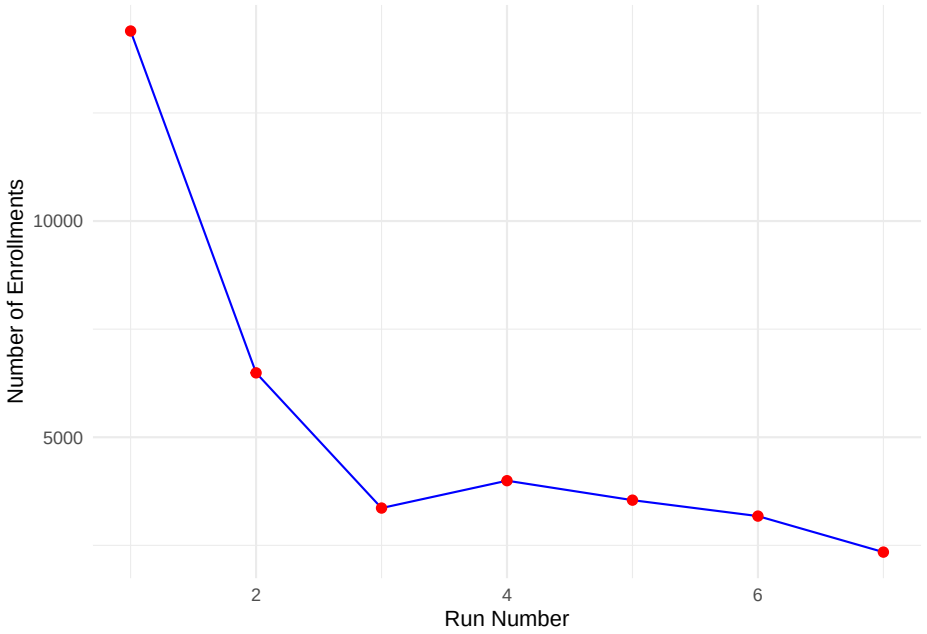

Figure 1: The enrollment trend over runs

The plot reveals a significant drop in enrollments after the first run, followed by a steady decline in subsequent runs, stabilizing by the seventh run. This trend suggests a high initial interest that rapidly declines, possibly indicating issues with course content, changing market dynamics, or competition. It may also reflect a saturation of the target audience.

Gender Equality Trend Analysis:

Stacked percentage bar charts and line plots to visualize gender distribution per run.

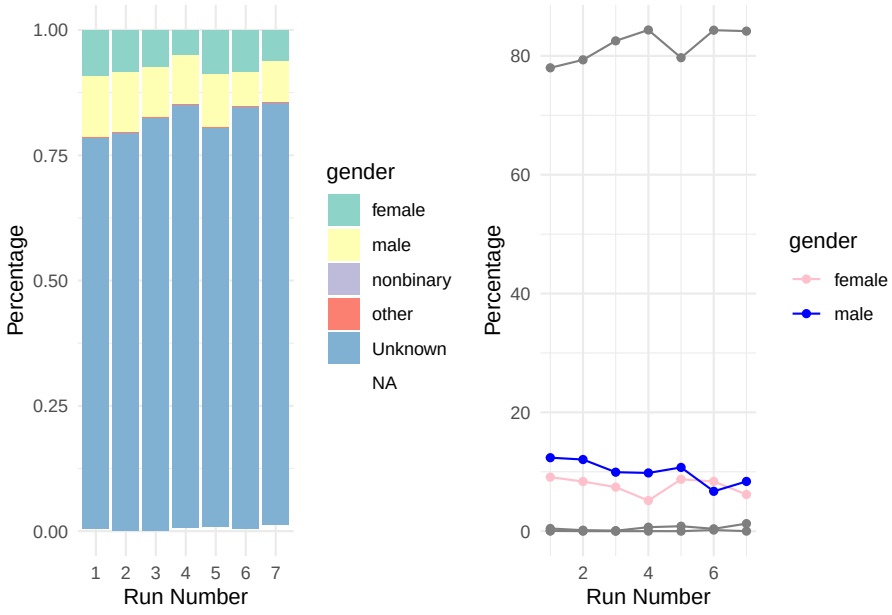

Figure 2: Left: Stacked gender percentage per run. Right: The trend of gender percentages over runs

Approximately 80% of gender data is marked as "unknown," which significantly impacts the reliability of gender-based insights. However, from the available data, female participation appears to start at 10%, dips to 5% around the fourth run, and then fluctuates, while male participation remains fairly consistent between 10-15%. Notably, female participation surpasses male in the sixth run.

The high volume of unknown gender data indicates a need for better data collection. The observed fluctuations in known gender participation could reflect changes in course appeal to different genders or broader societal trends.

Correct Response Trend Analysis:

Line plots and bar charts to show the trend of correct quiz responses.

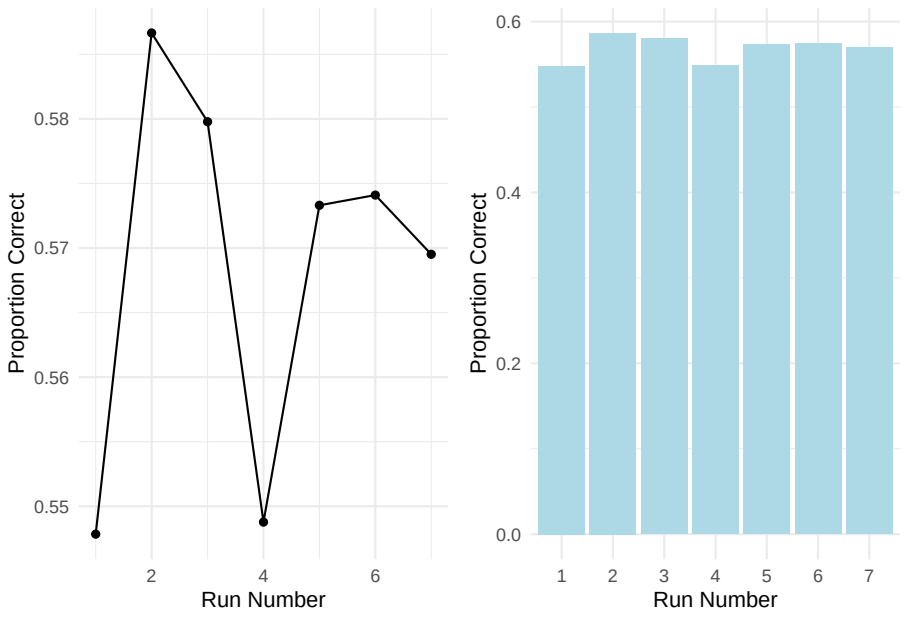

Figure 3: Left: Proportion of Correct Answer. Right: The tren of correct answers' proportion over runs

Here is a more comprehensive model demonstrating the distribution of the rate for correct responses and the mean:

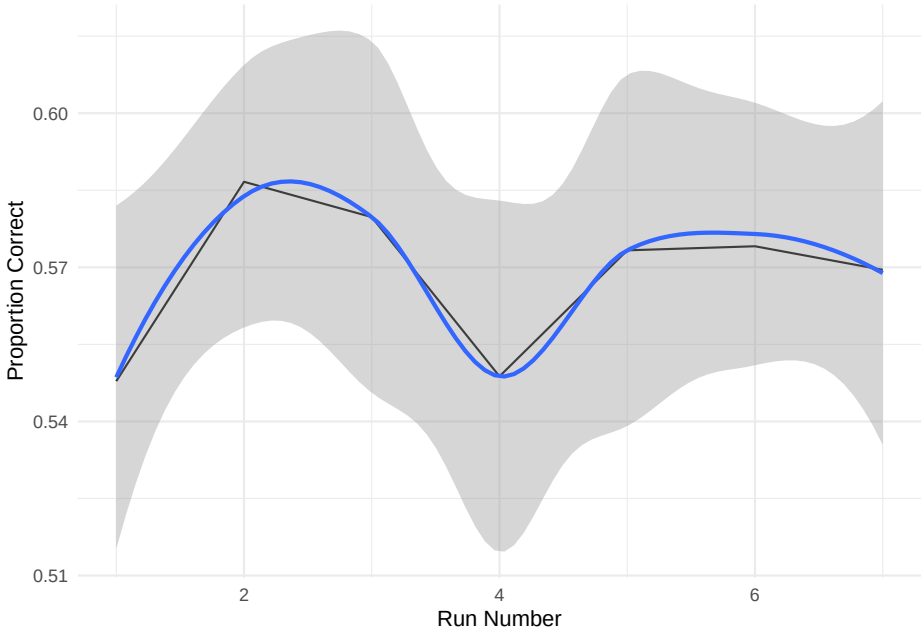

Figure 4: The trend of correct answers over run

The percentage of correct answers ranges between 51% to 61%, showing a slight increase in the second run, a dip in the fourth, and a general trend of stability with minor fluctuations.

The relative stability of correct responses suggests consistent course difficulty and learner understanding. The

fluctuations might be influenced by the changing composition of the course audience or minor modifications in course content or assessments.

Step Completion Time Trend Analysis:

Line plots depicting the average time taken for specific course steps across runs.

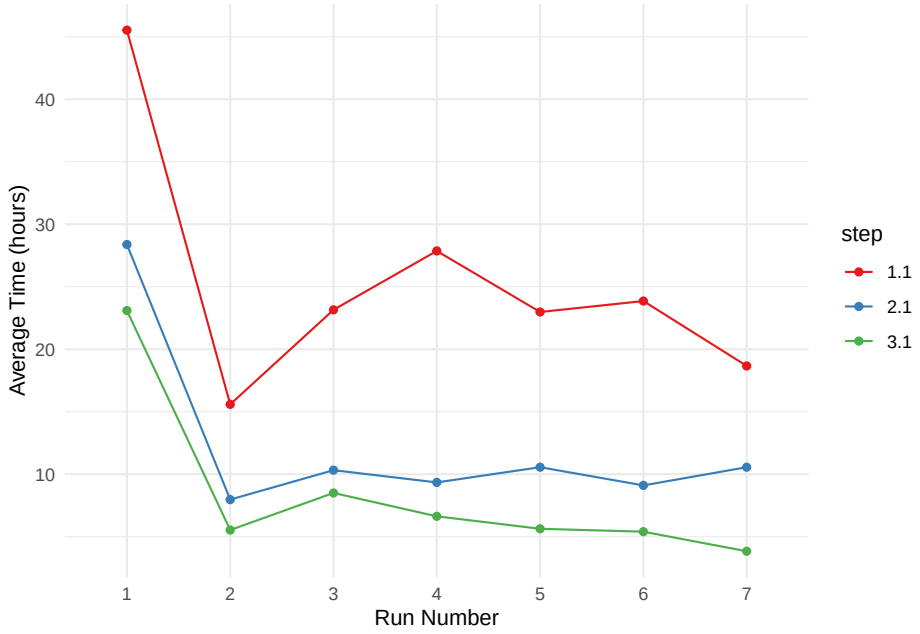

Figure 5: Average Time Taken for Steps '1.1', '2.1', and '3.1'

There's a noticeable decrease in the average time taken for step 1.1 from the first to the second run, with subsequent runs showing smaller fluctuations. Steps 2.1 and 3.1 show a similar but less pronounced trend.

The initial drop in time taken for step 1.1 might indicate a learning curve effect or adjustments in the course content making it easier to complete. The trends for steps 2.1 and 3.1 suggest a consistent engagement level across these course sections.

Completion Percentage Analysis:

Bar charts representing the completion percentage for each course step across runs.

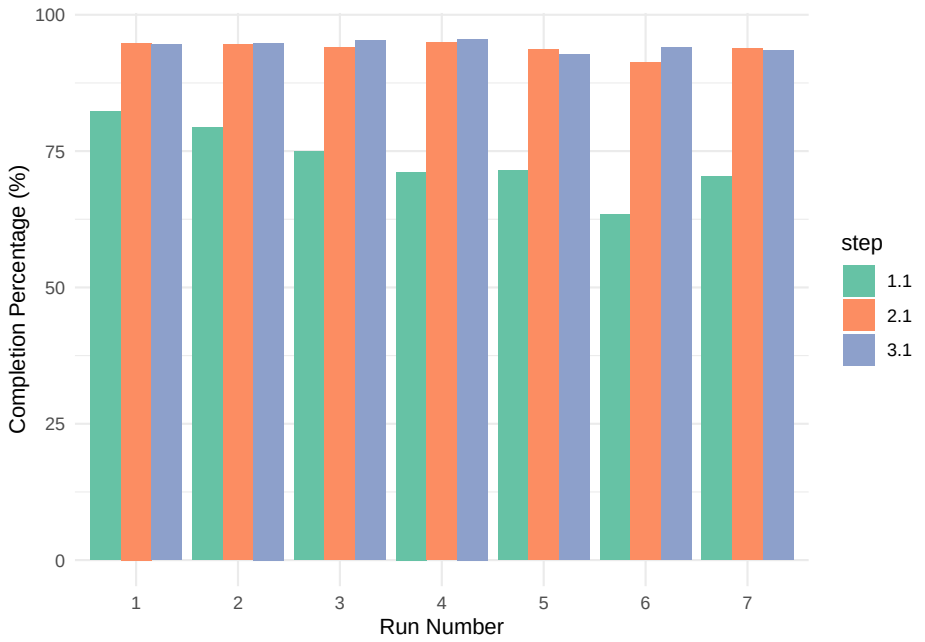

Figure 6: Completion Percentage for Steps '1.1', '2.1', and '3.1'

Step 1.1 shows a gradual decrease in completion percentage from 83% in the first run to 70% in later runs, while steps 2.1 and 3.1 maintain a high completion rate of around 90% across all runs.

The higher completion rates for steps 2.1 and 3.1 could imply these sections are more engaging or perceived as more valuable by learners. The decreasing trend in step 1.1 completion might point to issues in the initial course sections that could be leading to early drop-offs.

Evaluation

The data analysis of the MOOC "Cyber Security: Safety At Home, Online, and in Life" has unveiled several important trends and patterns. The enrollment trend shows a robust initial interest, followed by a gradual decline over subsequent runs, suggesting the course's appeal has diminished over time. This observation points to potential areas for improvement in course marketing, content updating, or adaptation to evolving learner preferences and market trends. Gender equality analysis, though limited by a significant amount of unknown gender data, reveals fluctuations in known gender participation. This highlights an opportunity for more focused outreach and content adjustments to enhance gender inclusivity.

In terms of learner performance and engagement, the course exhibits a general stability in correct response rates, suggesting consistent course difficulty and learner understanding. However, minor variations in this trend could be indicative of changing learner demographics or subtle shifts in course content, meriting further investigation. Analysis of step completion times and rates reveals insights into how learners interact with the course material. A decreasing trend in the time taken for initial steps and varying completion rates across different steps suggests that while the course retains engagement in its later sections, there may be issues with initial engagement. This observation is crucial for guiding improvements in the course's introductory content.

These findings lay the groundwork for the next phase of analysis: investigating specific learner feedback from archetype surveys, reasons for leaving the course, and weekly sentiment surveys. Understanding the learner's personality, motivations for course discontinuation, and weekly perceptions will provide a richer context for enhancing course quality. This next step is vital in pinpointing specific areas for improvement, aligning the course more closely with learner needs, and ultimately reversing the trend of declining enrollments.

Round 2 of the CRISP-DM Cycle

Business Understanding

In the first round of analysis, I observed significant trends in enrollment, gender equality, learner performance, and engagement with the course material. These insights align with my primary objective of deriving meaningful information for businesses in the online education sector. Building on these findings, the next research question focuses on a deeper aspect of learner engagement and course quality:

"Where can the course quality be improved based on learner feedback?"

This question is crucial as it delves into the qualitative feedback from learners, including their personal attributes from archetype surveys and reasons for course discontinuation, to identify specific areas for improvement. By focusing on this aspect, I aim to uncover actionable insights that can directly influence course enhancements, addressing the declining enrollment trends and varying engagement levels identified in the first round. My objective remains to provide businesses in the online education sector with valuable insights for optimizing their course offerings.

The overall scope of the analysis remains consistent with my initial objectives, now with a refined focus on leveraging qualitative feedback to drive improvements in course design and content.

Data Understanding

In this phase, I reassess my data in light of the newly defined research question. The primary data sources for this round include learner feedback from archetype surveys, reasons for course discontinuation, and weekly sentiment surveys. These sources are rich in qualitative information that can provide deeper insights into learner experiences and perceptions.

While the data quality from the first round was found to be sufficient, I now aim to extract more detailed insights from the qualitative feedback. This may involve employing text analysis techniques to analyze written feedback or thematic analysis to identify common patterns and themes in learner responses. The challenge lies in translating qualitative feedback into actionable data points that can inform specific improvements in course content and structure.

The datasets from the surveys are composed of:

- Archetype survey data (collected from the last 4 runs)
- Leaving survey data (collected from the last 4 runs)
- Sentiment survey data (collected from the last 4 runs)

Given the depth and relevance of the existing data, it is deemed fit for this second round of analysis. My next steps involve planning the extraction and analysis of qualitative feedback, which will be crucial in identifying specific areas for course improvement.

Data Preparation

In this second cycle of analysis, the data preparation phase focused on merging and processing survey data from the last four runs of the MOOC. This phase was critical for extracting actionable insights from learner feedback, aligning with my objective to identify specific areas for course improvement.

Select Data

In this phase, I concentrated on the last four runs of the course, primarily due to the availability of adequate and effective survey data from these iterations. This focus allows me to analyze the most recent and relevant feedback, directly aligning with the evolving nature of the course as observed in the first cycle of analysis, where trends in enrollment, gender distribution, and quiz performance were evaluated.

Integrate Data

I developed a merge_survey_data function to integrate data from archetype surveys, leaving surveys, and weekly sentiment surveys. This integration was essential in creating a comprehensive view of learner feedback across different dimensions.

A loop was used to merge survey datasets for each of the last four runs, with the id field standardized to ensure consistency across datasets. This process resulted in the creation of a combined dataset, combined_survey_data, which represented a unified view of learner feedback across these runs.

id
learner_id_archetype
responded_at_archetype
archetype
learner_id_leaving
left_at
leaving_reason
last_completed_step_at
last_completed_step
last_completed_week_number
last_completed_step_number
responded_at_weekly
week_number
experience_rating
reason
run_number

Construct Data

I constructed derived attributes that included proportions of different learner archetypes and aggregated reasons for leaving the course. These attributes are pivotal in understanding the diversity of the learner population and their motivations for disengagement. This approach allows me to correlate these qualitative insights with the engagement patterns and demographics observed in the first analysis cycle, bridging the gap between quantitative data and qualitative feedback.

Clean Data

The "id" field was standardized across datasets to ensure uniformity. This step was crucial to accurately merge and compare data across different surveys and runs. The function "check_and_convert_id" was employed to convert "id" to character strings where necessary.

Format Data

Formatting adjustments, like converting "run_number" to a factor, were made to prepare the data for analysis. This ensured consistency with the data structure used in the first cycle, facilitating comparative

analysis.

Dataset

The resultant data-set, "combined_survey_data", encompasses feedback from the recent runs, including learner archetypes, leaving reasons, and weekly sentiments. Text data was further processed for word cloud visualization, removing common and custom stop-words to focus on significant terms.

word	learned	information	lot	content	easy	informative
n	8	7	7	5	4	4

This data-set provides a qualitative dimension to the quantitative findings from the first cycle. For instance, analyzing the proportions of archetypes and reasons for leaving gives a deeper understanding of why learners may disengage from the course, and the word cloud visualization of weekly sentiment surveys highlights the key themes in learner feedback.

Modelling

In this phase of the analysis, I focused on visualizing and understanding the qualitative feedback data from the last four runs of the course. The objective was to uncover patterns in learner archetypes, reasons for leaving, and key themes in their weekly feedback.

To understand the diversity of learner types and their evolution across course runs, I created several visualizations of the archetype proportions:

Trend Analysis of the Proportions for Each Archetype within Each Run

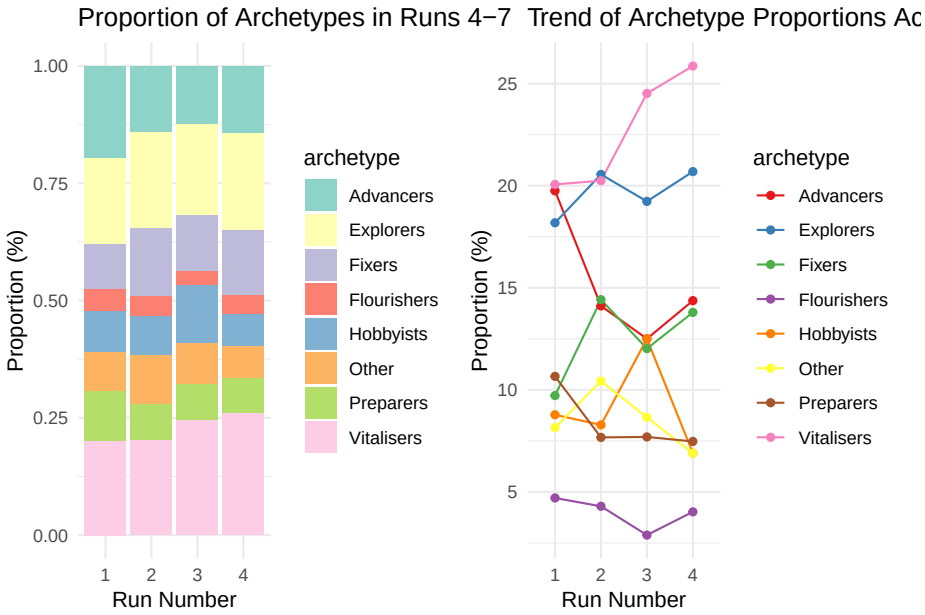

Figure 7: Left: Proportion of archetypes in runs 4-7. Right: Trend of archetype proportions across runs 4-7

Here I want to dive dipper in how the proportion of each archetype changes within each run:

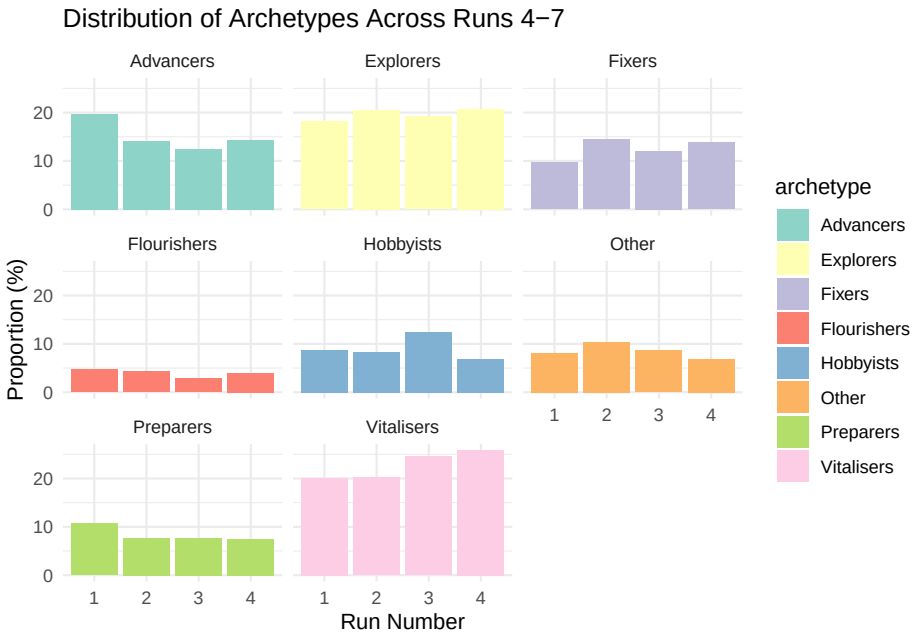

Figure 8: Distribution of archetypes across runs 4-7

From these visualizations, it was observed that certain archetypes like “Advancers” and “Preparers” showed a decrease over time, while “Vitalisers” increased. “Explorers”, “Fixers”, “Flourishers”, and “Hobbyists” remained relatively stable. These trends provide insights into how different learner types are interacting with the course over time.

The Summary of Leaving Reasons

Figure 9: Ranking of reasons for leaving the course

An ordered dot plot was used to rank the reasons for leaving the course. The most common reasons were lack of time and unspecified other reasons, followed by preferences not to say and misaligned expectations with the course. This ranking is vital in understanding the primary barriers to course completion.

The Word Cloud of User Experience Survey

Figure 10: The summary of leaving reasons

A word cloud from the weekly sentiment survey responses highlighted the most frequently used words by learners. Words like "Learned," "Information," and "Content" were prominent, indicating a positive reception of the course material.

Evaluation

In the second cycle, the focus on qualitative feedback brought to light nuanced aspects of learner engagement. The analysis of learner archetypes revealed distinct trends in learner diversity and preferences. Archetypes like "Vitalisers" saw an increase in representation, while "Advancers" and "Preparers" decreased over time. These shifts suggest changing learner demographics and needs, which are crucial for tailoring course content.

The reasons for leaving the course, such as time constraints and mismatched expectations, indicate key areas where the course could be improved. Understanding these factors is critical in addressing dropout rates and enhancing learner satisfaction. The word cloud from the sentiment surveys underscored a generally positive reception of the course content, with terms like "Learned," "Information," and "Content" being prominent.

Integrating from the First Cycle

These qualitative insights complement the quantitative findings from the first cycle. The declining enrollment trends from the first cycle can now be viewed through the lens of the changing archetype proportions and reasons for leaving observed in the second cycle. While initial interest in the course was high, sustaining engagement seems to be a challenge, likely influenced by the evolving learner profiles and their expectations.

The gender distribution analysis in the first cycle, despite the limitation of high unknown data, now gains additional context. The fluctuations in known gender participation could be linked to how different archetypes perceive and interact with the course, informing potential strategies for more inclusive content and marketing.

Final Conclusions

Combining insights from both cycles provides a comprehensive understanding of the course dynamics. The key takeaway is the importance of adapting to the changing needs and preferences of learners. Clear communication of course demands, adjustments in content to align with learner goals, and addressing specific barriers like time constraints are essential.

The course content is well-received, but the challenge lies in aligning it with the diverse and evolving learner base. By addressing the identified issues of engagement and expectation management, there's potential to reverse the declining enrollment trend.

In summary, this holistic analysis underscores the need for a dynamic, responsive approach in online course management, and insights reveal the course's success relies on adapting to my learners' evolving needs. Continual adaptation to learner feedback and market trends is vital for maintaining relevance and effectiveness in the rapidly evolving field of online education.

My analysis acknowledges the limitations inherent in my methodology and data. These include the potential impact of missing data, constraints in the generalisability of findings, and the assumptions underlying my statistical analyses. I believe that transparency in these areas strengthens the credibility of my conclusions.

Deployment

In this final phase of the CRISP-DM process, the primary goal is to effectively disseminate the findings from my comprehensive analysis of the MOOC "Cyber Security: Safety At Home, Online, and in Life." To achieve this, I have developed a detailed report and an accompanying presentation, tailored to convey the key insights to stakeholders involved in the course's development and management. These deliverables highlight critical findings such as enrollment trends, learner archetype dynamics, reasons for course discontinuation, and key themes from sentiment analyses.

The report and presentation are designed to provide a clear narrative of the course's performance, showcasing both quantitative and qualitative insights. The visualization of data, including trends in enrollment, graphical representations of learner archetypes, and sentiment analysis word clouds, ensures that the information is easily interpretable and actionable. Focusing on these elements aims to assist stakeholders in understanding the underlying patterns and taking informed steps towards enhancing the course.

To ensure the long-term impact of these findings, a monitoring and improvement plan is proposed. This includes regular reassessment of course data and continuous feedback analysis to adapt to evolving learner needs and market trends. The plan emphasises the need for iterative improvements to course content, marketing strategies, and learner support, informed by the data analysis.

Reflecting on the project, it becomes clear that a structured, data-driven approach is essential in understanding and improving online learning experiences. The thorough documentation of each phase of the analysis, from data preparation to modelling and evaluation, not only serves as a comprehensive record of the work done but also as a valuable guide for me to do any future endeavours in this field. This structured approach, combined with the clear communication of findings, ensures that the project's insights will effectively guide the course's evolution and improvement.