Cyber Security MOOC Data Analytics

Rushi Girdharbhai Vasoya

2026-01-16

CRISP-DM Cycle 1: Understanding Learn Engagement Across Course Runs

Business Understanding

Online learning platform, like Newcastle University, Collectively produce hundreds of thousands if not millions of record of learners interaction across multiple run of course. Course teams and the platform analysis seek insight into how learners engage with the course content as a part of course effectiveness and future design decision. Specially, While environment figures may indicate the reach of a course they do not necessarily reflect the way in which learner actively participate once enrolled.

The first investigation here in explores general trends in the pattern of learner engagement across multiple runs of course. Engagement is considered at the foundations level using enrollment data and the step activity records, which together provide insight into both learner participants and interaction with the course content. This foundation understanding must be established before more complex behavior indicators can be introduced in later analysis.

This study will try to establish whether engagement pattern are the same between run and explorer to what extent enrollment is covered into actual participation. The study is deemed to be successful if through the use of descriptive Statistics and visualization pattern of engagement can be described clearly and these findings set a meaningful foundation for future investigation in cycle 2.

Data Understandig and Preoaration

To data analysis relies on two main data sources provided from the Newcastle University: Enrollment data and step activity data. The enrolment data tracks when learner enrolled for each run of our course, while step activity record how learners have engaged with specific steps of Course.

Data from serveral runs of course where aggregated into structured data set. This data set represent a number of rows with several thousand entries from learner enrollment and intractions at a step level, with columns having learner IDS, Run IDS, steps and activity value the data column includes both category and numerical variables.

Initial discovery yielded a number of issues that impact data quality. A portion of student are represented in enrol data without being reflected in activity data, indicating that they do not participate after enrolling. Also, activity levels are highly variable from run to run, which is in a variability in activity distributions. This have been maintained since these are real activity patterns in learners.

Only enrollment and step activity data were chosen for this cycle in order to keep the focus of the analysis clear and easy to interpret. Both of these datasets are the most basic form of indicators for engagement and can be easily compared at highly level without adding any complexity of activity. The video and response data will be saved for investigation two.

For the diagnosis the environment file as well the step activity file from all the runs were combined in a consolidated form. The identifier for the running of the courses were maintained in order to allow further comparison of data, while aggregation was done in order to obtain the total of the enrollments and the step activities for every run.

Data Analysis

This section addresses the research question: what are overall patterns of the learner engagement across all course runs? The analysis is bounded by descriptive Comparisons of learner enrollment and step activity across multiple runs of the course in order to find broad engagement trends.
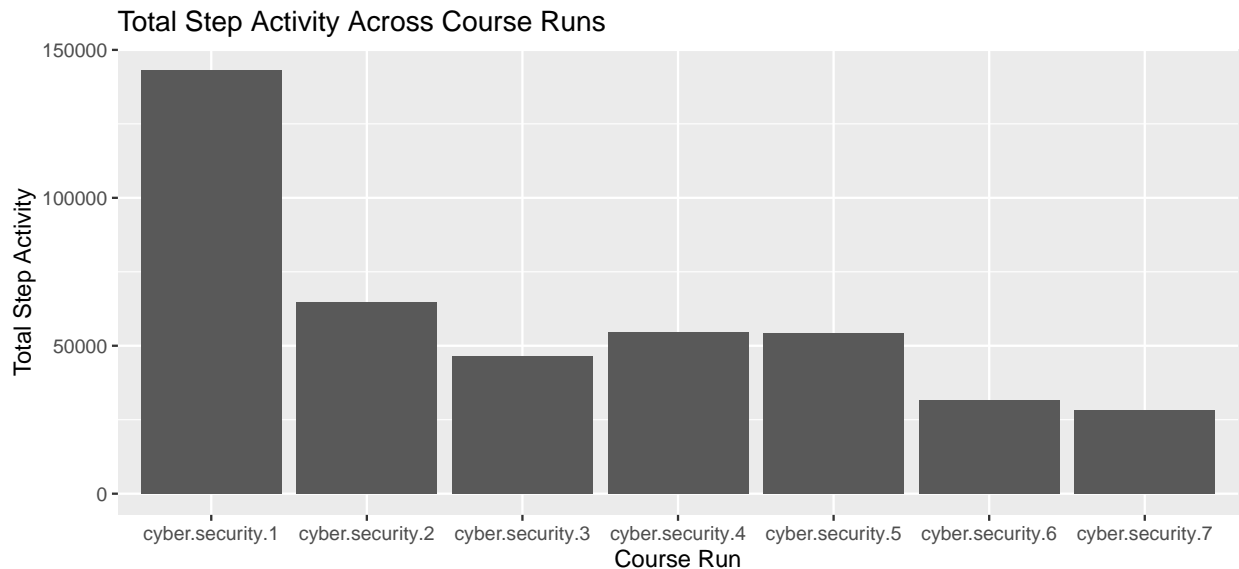
Learner engagement is considered from two complementary viewpoints. First, enrollment data is used to indicate the level of learner participation At a point of registration. Second step activity data is used to represent observable engagement with course content after enrollment. Comparing these measure occurs runs allow the variation in engagement to the be assessed and whether enrollment Translates into active Participation.

Descriptive statistics of total enrollments and total step activity were generated for each course run. The result have allowed the comparisons of engagement pattern across different iterations of course at a high level.

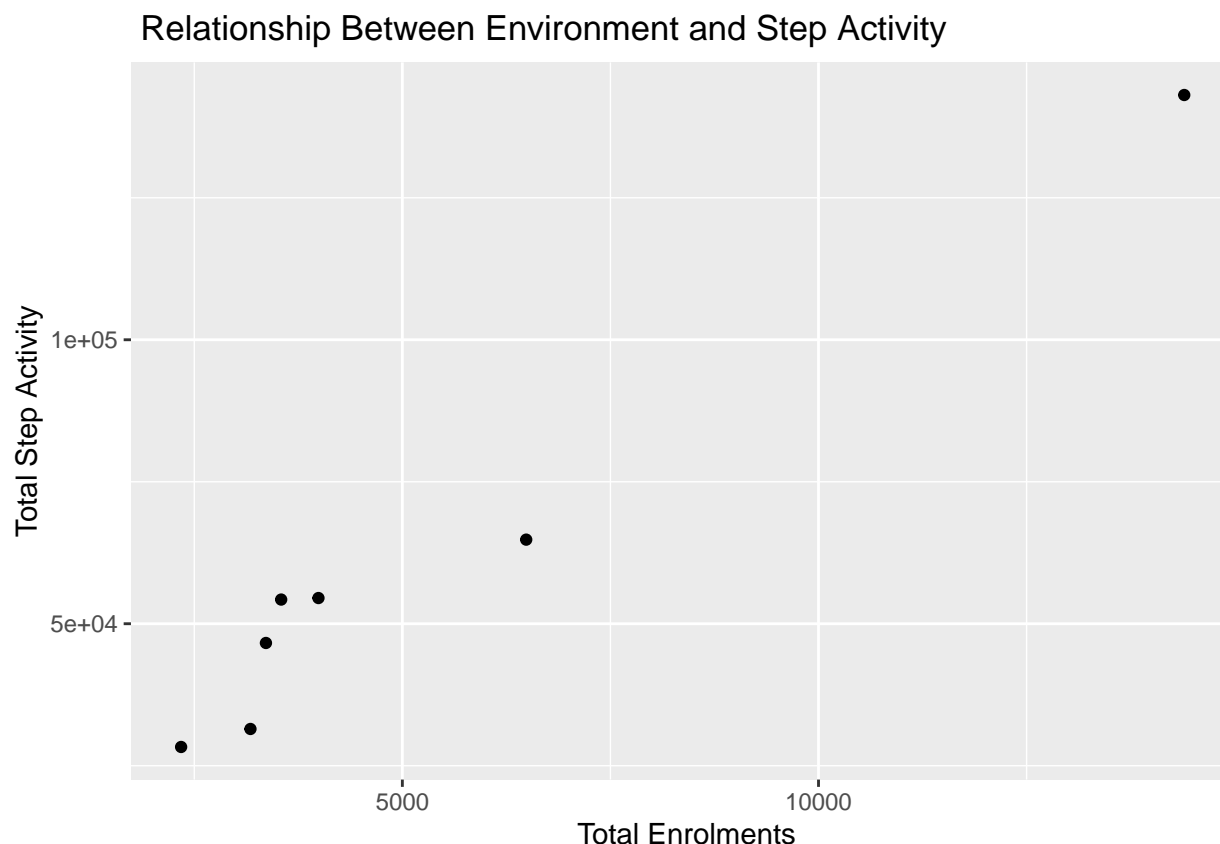Table 1: Summary of enrollments and step activity occurs course runs

run_id	total_enrolments	total_step_activity
cyber.security.1	14394	143092
cyber.security.2	6488	64809
cyber.security.3	3361	46614
cyber.security.4	3992	54524
cyber.security.5	3544	54257
cyber.security.6	3175	31472
cyber.security.7	2342	28304

To further support interpretation, graphical representation Were employed to illustrate contrast in the level of engagement for multiple instance of a course.

Analysis of the findings indicates that the level of engagement of learners differ significantly across the runs of the courses. Although some courses have high level of step engagement, other tend to have low levels of engagement, yet the number of enrolled learners in the two scenarios is similar. This implies that the behavior of learners in terms of their engagement with the course differs across the runs of the courses.

Further analysis of the relationship between enrollments and step activity data reveals that greater enrillments are not necessarily correlated to greater engagement levels.

Relationship Between Environment and Step Activity

In other runs, There may be many learners who enroll, but there may be fewer engagement with course steps. In this case It clearly show that there are learners who become disengaged after enrolling making it crucial to include behavioral engagement measures in addition to participation counts.

Overall, it can be seen that there is a descriptive analytical summary about engagement that proves engagement intensity is not consistent between runs and does not scale with enrollment.

Evaluation

Objective of cycle one was to build a founded fundamental understanding of learner engagement by leveraging enrollment data and data from step activities. Investigation cycle 1 was successfully in establishing pattern of learner engagement and the fact that learner participation and engagement differ from each other.

These results satisfy requirements for success that were identified in business understanding phase. Data from multiple course runs has been combined successfully, and engagement has been summarized as both numerical and graphical it is evident that data on environments is not adequate in describing engagement and the data on step activity is useful in Complementing data environments.

These outcomes highlights some critical Limitations in measure of high level engagement. It can be noted that step activity provides insight into learner engagement volume, Though it fails to provide an understanding

of learner activity in terms of different categorical of content. Consequently, these outcome generate further the concern in relation to learner engagement activity.

Therefore, findings from cycle 1 provide a direct motivation for what is explored cycle 2. In the subsequent study, a deeper understanding of learner engagement will be gained by including a far Sophisticated Set of data points related to learner engagement, like video based, interaction based, learning response, and architect data.

CRISP-DM Reflection(cycle 1)

This investigation follows the success of CRISP-DM in going from business understanding to data understanding, data preparation, analysis, and final evaluation. Cycle one takes an exhaustive and closed investigation approach to create descriptive baseline measure of learner engagement over various course deliveries. The finding from this cycle will thus from an effective foundation of for cycle 2, where There will be an analysis of specific behavior data to justify learner engagement behaviors.

CRISP-DM Cycle 2: Behavioural Engagement Analysis Across Course Runs

Business Understanding

Cycle 1 made clear that learner engagement levels vary quite significantly across course runs, and high enrollment does not necessarily equate to high engagement. These findings were preliminary in nature because they were based on very high-level metrics of participation and didn't explain why these patterns occur.

The focus of Cycle 2 is the explanation of the engagement patterns identified in Cycle 1 through the analysis of learner interaction behaviors and learner archetypes. Rather than introducing new datasets, this cycle focuses on the use of most relevant behavioral and segmentation data in order to provide explanatory insight.

Behavioral engagement is analyzed using video interaction data and question response activity; learner archetypes are used to segment learners into meaningful groups, allowing differences in engagement to be interpreted in terms of the ways in which learners interact with course content and types of learners who are more or less engaged.

The investigation will be successful if clear explanation of variation in engagement observed in Cycle 1 is eminently done through patterns of behavioral engagement and archetype-based differences.

Data Understanding and Preparation

Cycle 2 depends on three main datasets that were provided by the Newcastle University MOOC platform:

- Video Stats CSV: Records learner interaction with video-based course content
- Question Response CSV: recording active leaner engagement through assessment interaction
- Archetype CSV: providing learner classification for behaviour segmentation.

All datasets were cleaned and combined across different course runs. A consistent run_id identifier was maintained to support comparison across course runs.

Question response data were summarized at the learner–run level to measure active participation behavior. Learner archetypes were then attached to these summaries in order to enable segmentation analysis. Video interaction data were combined across runs to represent depth of content engagement.

Data sets included only those that were necessary to explain Cycle 1 engagement patterns so that the analyses would be focused, interpretable, and methodologically sound.

Data Analysis

This section addresses the Cycle 2 business question: How do learner interaction behaviors and learner archetypes explain the engagement patterns identified in CRISP-DM Cycle 1?

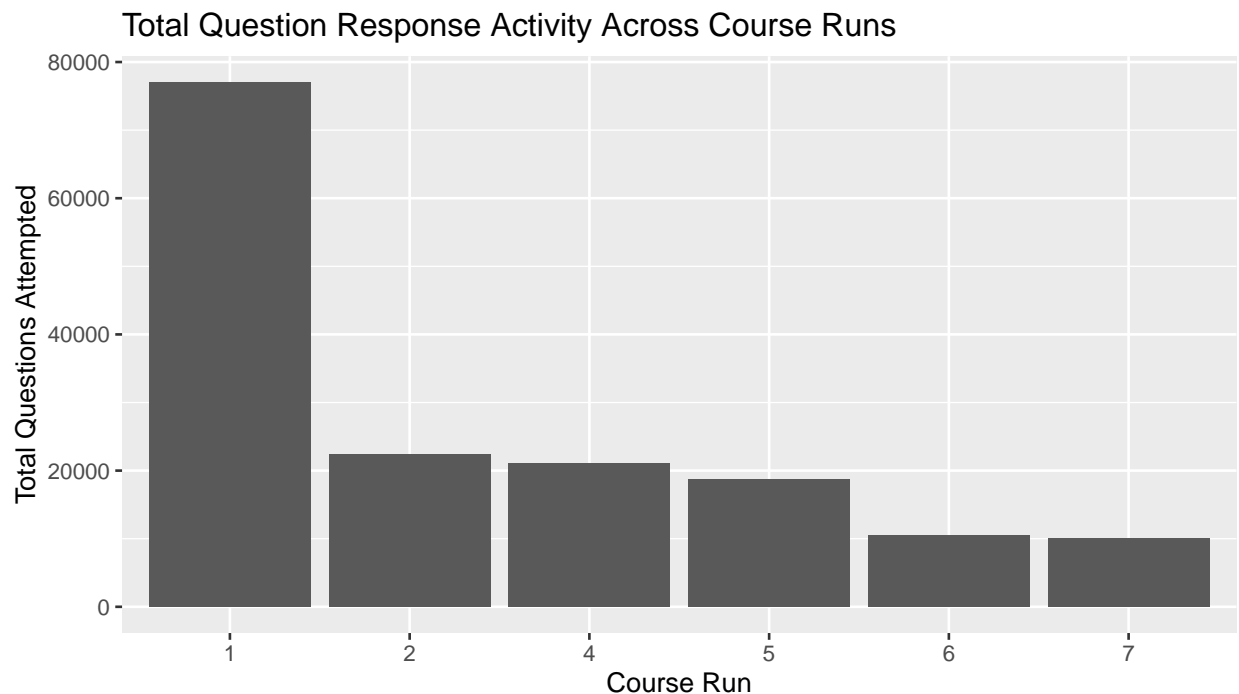** Behavioural Engagement: Question Response Activity **

Active participation was first explored by means of learner question response behaviour. Learner level summaries capture the number of questions attempted and correct per course run.

Table 2: Summary of Question Response Behaviour Across Course Runs

run_id	learners	total_questions_attempted	total_questions_correct
1	3410	77002	0
2	1379	22463	0
4	1148	21116	0
5	1131	18752	0
6	690	10533	0
7	623	10077	0

The above summary indicates that the levels of active participation differ from one course run to another, thus supporting the results from Cycle 1 that the intensity of participation can be different from run to run when the levels of enrollment are relatively similar.

In order to conceptualise this pattern, graphing the question response activity against the course runs was used.

** Behavioural Engagement: Video Interaction Activity **

Depth of engagement was further assessed using data related to video interaction. Activity records in videos represent learning engagement beyond enrolled activity.

Table 3: Summary of Video Interaction Activity Across Course Runs

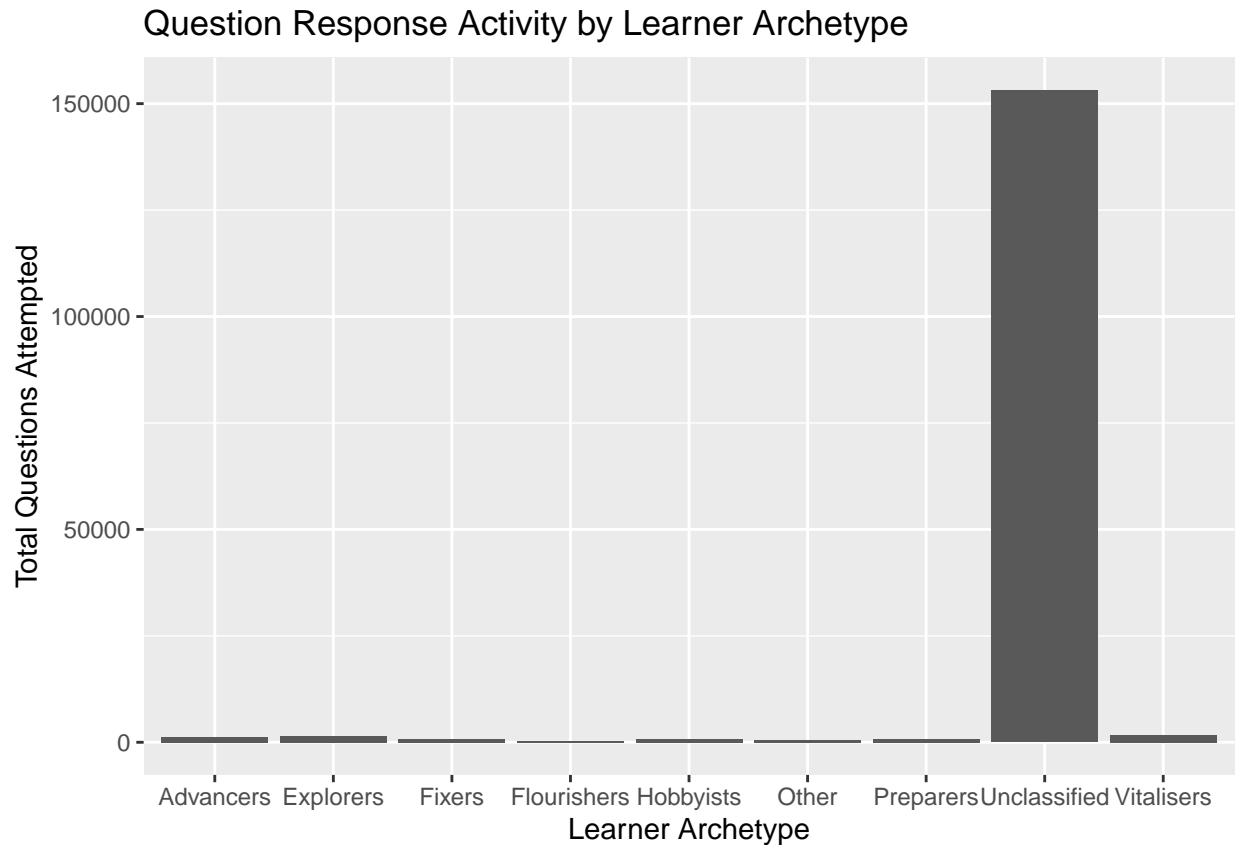| run_id	total_video_activity
3	13
4	13
5	13
6	13
7	13

These results show that there are course runs that have relatively high levels of video interaction yet relatively low levels of active participation, and vice versa. It is this contrast that has accounted for the inconsistencies shown for engagement, as mentioned in Cycle 1, and that equally, different types of behavioural engagement do not always show consistency.

** Learner Archetypes and Engagement Behaviour **

In order for the variations in the level of engagement to be understood, the behavior of the learners was segmented based on archetype categories.

Table 4: Question Response Behaviour by Learner Archetype

archetype	learners	total_questions_attempted	total_questions_correct
Advancers	60	1175	0
Explorers	66	1269	0
Fixers	36	648	0
Flourishers	13	241	0
Hobbyists	37	762	0
Other	34	527	0
Preparers	30	666	0
Unclassified	7866	153128	0
Vitalisers	82	1527	0

Question Response Activity by Learner Archetype

A bar chart titled "Question Response Activity by Learner Archetype" with the y-axis labeled "Total Questions Attempted" (ranging from 0 to 150000) and the x-axis labeled "Learner Archetype". The x-axis categories are: Advancers, Explorers, Fixers, Flourishers, Hobbyists, Other, Preparers, Unclassified, and Vitalisers. The "Unclassified" category shows a bar reaching approximately 153000, while all other categories show values near 0.

An analysis of the data reveals the extent to which learner archetypes display varied behaviors concerning their level of engagement. Some archetypes play a pivotal role in being actively involved, whereas others remain less involved once they enroll. This explains the rationale behind the high number of enrollments not necessarily being reflected by a high level of engagement, as was witnessed during Cycle 1.

Evaluation

The aim for Cycle 2 involved the use of learner interaction behaviors and the use of learner archetypes to explain the learning engagement patterns that emerged during Cycle 1. This aim has been achieved.

The results show that variability in engagement levels for multiple runs of a course can be accounted for on the basis of variability in behavior interaction and learner archetype composition. Engagement of learners can be impacted not only based on volume of participation but also based on type and behavior of participants.

CRISP-DM Reflection (Cycle

Cycle 2 further expands on the lessons from Cycle 1 with the transition from describing participation measures to explaining through behavioral analytics. Once more, the process of following the CRISP-DM methodology helped in conducting a targeted and interpretable analysis.

Through the strategic use of behavioral interaction data and archetypes, it is clear from this cycle that engagement is complex and emerges in relation to behavior and type. The results here pose a compelling solution to the engagement issues presented in Cycle 1 and provide a great basis to move forward with in future analyzes using predictive models or outcome assessments.