MAS8600 Cyber Security MOOC Engagement — CRISP-DM Report-Two Cycles

Rishabh Rajan Yadav

12th January, 2026

Contents

Introduction **2**

CRISP-DM Cycle 1 — Engagement & Disengagement **2**

Business Understanding . . . 2
Stakeholders . . . 2
Stakeholder Goals . . . 2
Success Standards . . . 3
Data Understanding . . . 3
Data Sources . . . 3
Important Factors . . . 3
Surface Characteristics of the Information . . . 3
Considerations for Data Quality . . . 3
Data Preparation . . . 3
Data Analysis . . . 4
Completion Rate Profiles by Run . . . 7
Distribution of Completion Rates by Run (Boxplot) . . . 8
Evaluation . . . 9
Deployment . . . 10

CRISP-DM Cycle 2 — Assessment Engagement & Persistence **10**

Business Understanding . . . 10
Data Preparation . . . 10
Modelling . . . 11
Correlation Analysis . . . 11
Deployment . . . 12
Limitations . . . 12

Future Work 13
Final Conclusions 13

```
library(ProjectTemplate)

invisible(capture.output(
suppressMessages(
suppressWarnings(
ProjectTemplate::load.project(project.dir = knitr::opts_knit$get("root.dir"))
)
)
))
```

Introduction

Massive Open Online Courses (MOOCs), which provide students with flexible access to course material regardless of their location, have become a popular method for delivering education at scale. MOOCs have helped increase educational engagement, but they are also frequently linked to high learner turnover rates. Individual learning outcomes and the overall efficacy of these platforms are limited since a significant percentage of enrolled students drop out before finishing a course. This report analyzes learner engagement data from a Cyber Security MOOC that was offered over seven different course runs (Runs 1–7) using the Cross-Industry Standard Process for Data Mining (CRISP-DM). The analysis aims to comprehend how students connect with the course material over time, pinpoint the times when disengagement is most likely to happen, and investigate if participation in course assessments is linked to ongoing engagement. Two CRISP-DM cycles serve as the framework for the study. Finding patterns of engagement and disengagement across specific course steps is the main goal of the first cycle. The second cycle looks at the relationship between student engagement with quizzes and course persistence. ProjectTemplate is used to impose an organized workflow, while renv is used to guarantee package and environment consistency. All analysis is totally reproducible.

CRISP-DM Cycle 1 — Engagement & Disengagement

Business Understanding

Stakeholders

Many stakeholder groups involved in the development and provision of online education can benefit from the analysis's conclusions. These include curriculum designers who are in charge of creating learning pathways, online learning platform providers who aim to increase learner retention and completion rates, and course instructors who are in charge of teaching and assessment.

Stakeholder Goals

High learner dropout rates diminish MOOCs' instructional value and may jeopardize their long-term viability. Disengagement may be a sign that the material is confusing, excessively difficult, or badly organized from a teaching standpoint. Persistent dropout trends may indicate systemic problems with the learning process for platform providers. Stakeholders can prioritize changes such as modifying specific steps, adjusting course pacing, or adding targeted support at critical stages of the course by identifying where disengagement occurs.

Success Standards

This CRISP-DM cycle is considered successful if the analysis can clearly identify course steps associated with high disengagement, compare engagement patterns across multiple runs, and determine whether disengagement reflects consistent structural issues or varies across learner cohorts.

Data Understanding

Data Sources

For each of the seven course runs, two main datasets were used. Learner interactions with specific course steps, including whether a step was visited and whether it was completed, are documented in the step activity dataset. Learner interaction with quizzes, including submission behaviour and step-level accuracy, is captured by the question response dataset.

Important Factors

Learner identifiers, week and step numbers, timestamps recording first and last interactions, and indicators of quiz correctness are among the variables consistently used across both datasets. These variables enable analysis at both the learner and step levels.

Surface Characteristics of the Information

The datasets contain several thousand learners and several hundred individual steps across the seven runs. Although the completeness of interaction data varies between learners and runs, all interactions are timestamped. The open nature of the course and diversity of learner behaviour are reflected in the substantial variation in engagement intensity across cohorts.

Considerations for Data Quality

Initial data inspection revealed several quality issues typical of large-scale online learning datasets. These included irregular engagement patterns across runs, learners visiting steps without completing them, and missing or inconsistent timestamps. As the course was self-paced, such issues were expected and were addressed during data preparation to ensure the validity of subsequent analysis.

Data Preparation

All preprocessing was conducted using ProjectTemplate within the munge/ directory, ensuring that data cleaning and transformation procedures were applied consistently and reproducibly. Key preparation steps included standardising learner identifiers across datasets, parsing and cleaning timestamp variables, and creating indicators of learner progression and step completion.

Invalid records, including those with negative time differences caused by incorrect timestamps, were removed. Learner-level summaries capturing overall engagement and step-level funnel tables describing progression through the course were created for each run. No manual data loading or processing was used, preserving full workflow reproducibility.

Data Analysis

This cycle focused on understanding how learners progress through the course and identifying where engagement begins to decline. Engagement was treated as a sequential process rather than a single completion outcome, with learners moving through an ordered sequence of course steps.

Learner activity was summarised at the step level for each run, identifying how many learners visited, completed, and progressed beyond each step. From these summaries, step completion rates and step-to-step drop-off rates were calculated. This approach made it possible to distinguish between steps that attract initial interest but are rarely completed and those that learners are unlikely to continue beyond.

Proportional metrics were used instead of absolute counts to enable meaningful comparisons across runs with different cohort sizes. Engagement measures were therefore interpreted in relative terms, allowing consistent comparison of disengagement patterns across Runs 1 to 7.

Each run was analysed independently before cross-run comparisons were performed. This enabled the identification of disengagement patterns unique to specific cohorts as well as steps that consistently exhibited high drop-off across multiple runs. While isolated patterns were interpreted as cohort-specific variation, repeated drop-off at the same steps across several runs was treated as evidence of potential structural issues within the course.

Quantitative findings were interpreted using visual summaries. Funnel-style visualisations illustrated learner progression through the course, while completion-rate profiles highlighted steps associated with sharp declines in engagement. These visualisations supported interpretation and complemented the numerical analysis.

Disengagement was interpreted in context throughout the analysis. Although some attrition is expected in open online courses, very sharp drops were treated as potential indicators of issues such as increased cognitive load, unclear instructional design, or technical difficulties. The findings from this cycle form the foundation for the next CRISP-DM cycle, which examines learner persistence in relation to assessment engagement. ### Summary Statistics (Cycle 1)

```
funnel_step_all <- dplyr::bind_rows(lapply(1:7, function(r) {
get(paste0("funnel_step_run", r)) |>
dplyr::mutate(run = r)
}))
funnel_run <- funnel_step_all |>
dplyr::group_by(run) |>
dplyr::summarise(
total_step_visits = sum(visited, na.rm = TRUE),
mean_completion_rate = mean(completion_rate, na.rm = TRUE),
biggest_drop = ifelse(all(is.na(drop_off_to_next)),
NA_real_,
max(drop_off_to_next, na.rm = TRUE)),
.groups = "drop"
)
knitr::kable(
funnel_run,
caption = "Run-level summary of engagement metrics (Cycle 1)."
)
```

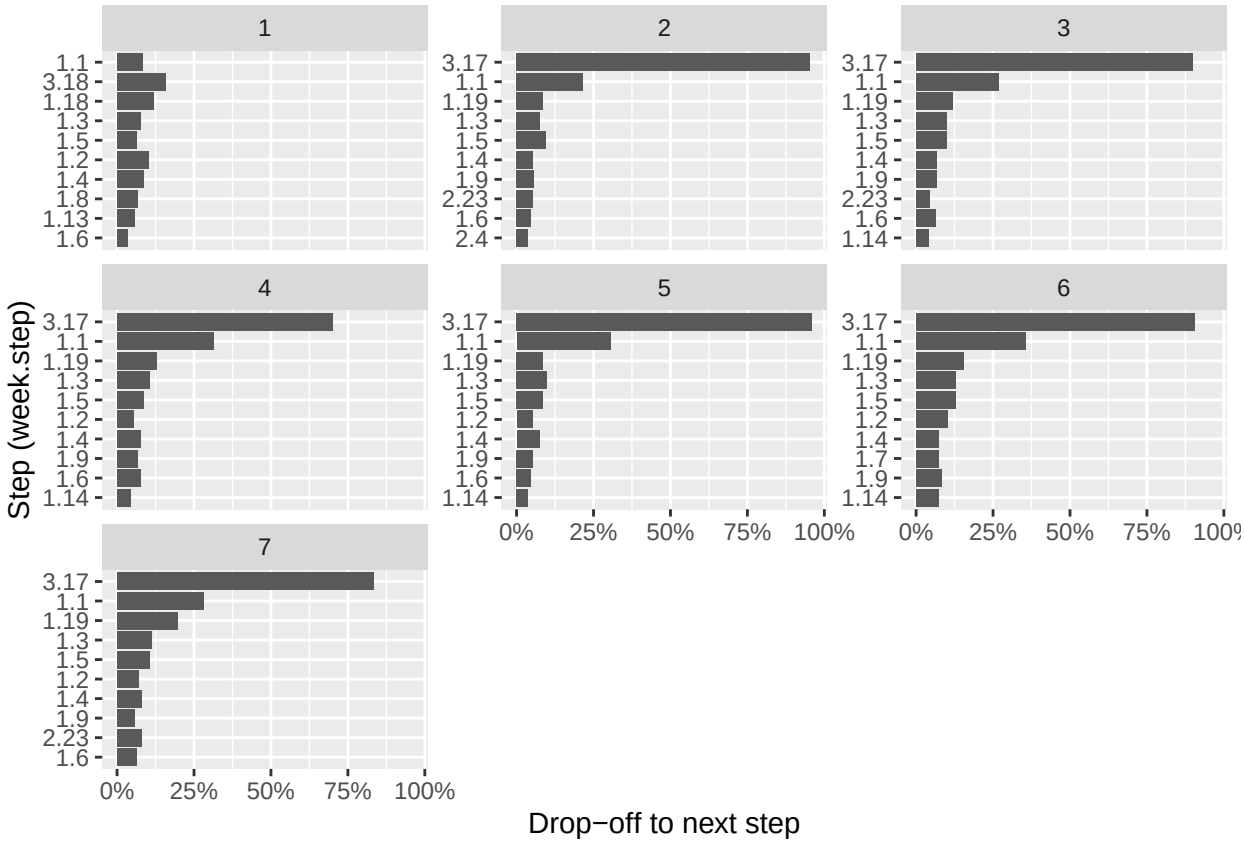
Table 1: Run-level summary of engagement metrics (Cycle 1).

run	total_step_visits	mean_completion_rate	biggest_drop
1	143092	0.9385613	0.1575630
2	64809	0.9338104	0.9534207

run	total_step_visits	mean_completion_rate	biggest_drop
3	46614	0.9285787	0.8987603
4	54524	0.9300287	0.7003367
5	54257	0.9137169	0.9601911
6	31472	0.9002498	0.9041096
7	28304	0.9101303	0.8321918

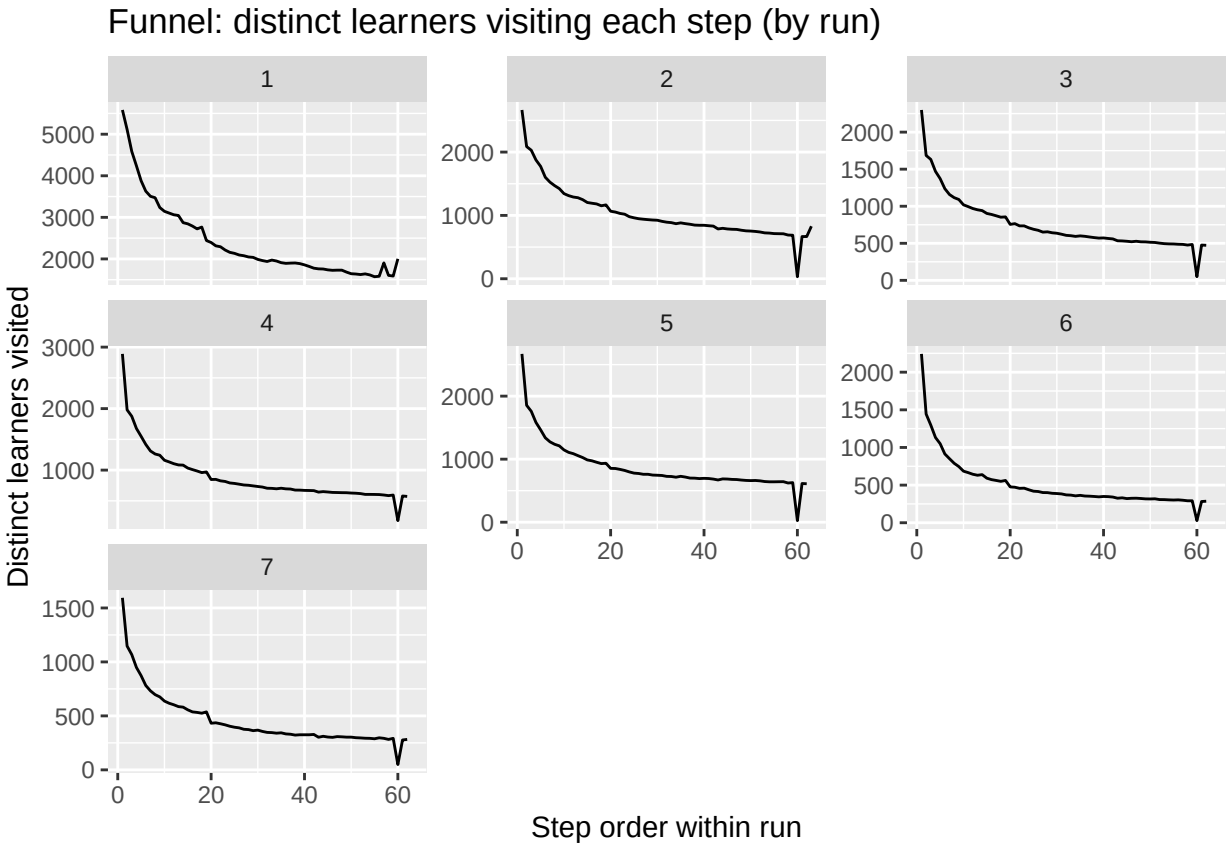
The largest observed step-to-step drop-off across all runs is r scales::percent(max(funnel_run$biggest_drop, na.rm = TRUE)). ### Modelling ### Identification of Largest Drop-Off Points

```
top_dropoffs <- funnel_step_all |>
dplyr::filter(!is.na(drop_off_to_next)) |>
dplyr::group_by(run) |>
dplyr::slice_max(drop_off_to_next, n = 10, with_ties = FALSE) |>
dplyr::ungroup() |>
dplyr::mutate(step_label = paste0(week_number, ".", step_number))
ggplot2::ggplot(
top_dropoffs,
ggplot2::aes(x = reorder(step_label, drop_off_to_next),
y = drop_off_to_next)
) +
ggplot2::geom_col() +
ggplot2::facet_wrap(~ run, scales = "free_y") +
ggplot2::scale_y_continuous(labels = scales::percent) +
ggplot2::coord_flip() +
ggplot2::labs(x = "Step (week.step)", y = "Drop-off to next step")
```



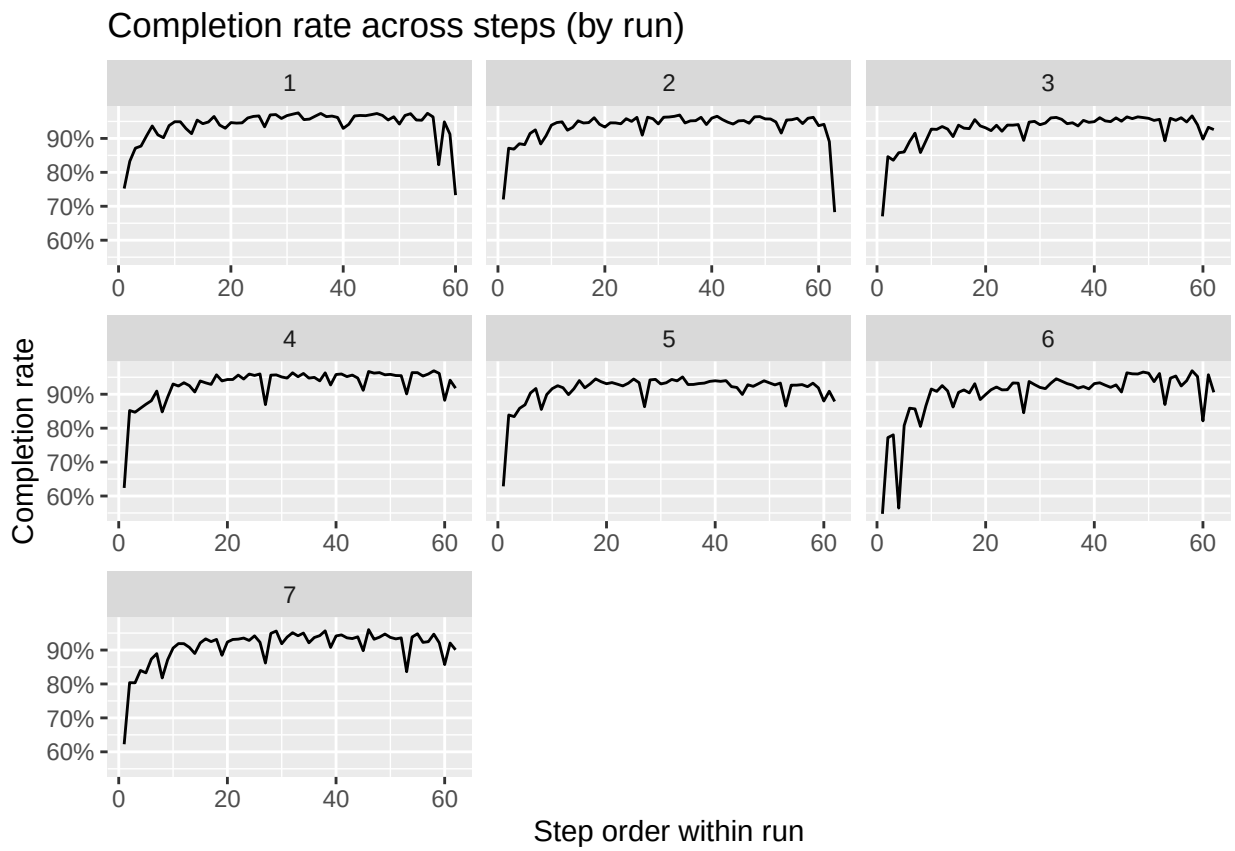
Funnel: Distinct Learners Visiting Each Step

```
stopifnot(exists("funnel_step_all"))
funnel_visits_plot <- funnel_step_all |>
dplyr::arrange(run, week_number, step_number) |>
dplyr::group_by(run) |>
dplyr::mutate(step_index = dplyr::row_number()) |>
dplyr::ungroup()
p_funnel_visits <- ggplot2::ggplot(
funnel_visits_plot,
ggplot2::aes(x = step_index, y = visited, group = factor(run))
) +
ggplot2::geom_line() +
ggplot2::facet_wrap(~ run, scales = "free_y") +
ggplot2::labs(
title = "Funnel: distinct learners visiting each step (by run)",
x = "Step order within run",
y = "Distinct learners visited"
)
p_funnel_visits
```



Completion Rate Profiles by Run

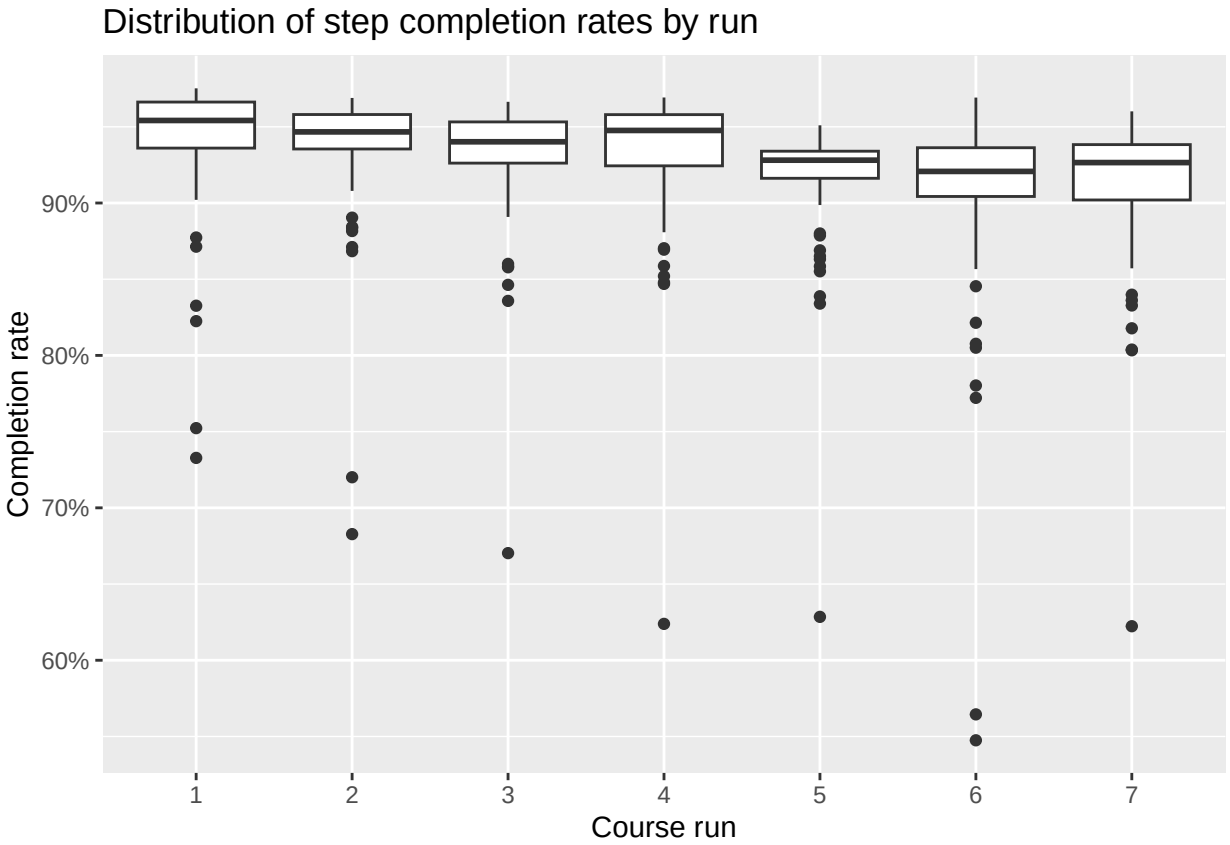
```
stopifnot(exists("funnel_step_all"))
funnel_step_plot <- funnel_step_all |>
dplyr::arrange(run, week_number, step_number) |>
dplyr::group_by(run) |>
dplyr::mutate(step_index = dplyr::row_number()) |>
dplyr::ungroup()
p_comp <- ggplot2::ggplot(
funnel_step_plot,
ggplot2::aes(x = step_index, y = completion_rate, group = factor(run))
) +
ggplot2::geom_line() +
ggplot2::facet_wrap(~ run, scales = "free_x") +
ggplot2::scale_y_continuous(labels = scales::percent) +
ggplot2::labs(
x = "Step order within run",
y = "Completion rate",
title = "Completion rate across steps (by run)"
)
p_comp
```



Distribution of Completion Rates by Run (Boxplot)

This boxplot summarises between-run differences in typical step completion rates and variability, complementing the step-order profiles shown above.

```
stopifnot(exists("funnel_step_all"))
p_completion_box <- ggplot2::ggplot(
funnel_step_all,
ggplot2::aes(x = factor(run), y = completion_rate)
) +
ggplot2::geom_boxplot() +
ggplot2::scale_y_continuous(labels = scales::percent) +
ggplot2::labs(
title = "Distribution of step completion rates by run",
x = "Course run",
y = "Completion rate"
)
p_completion_box
```



Evaluation Cycle 1 effectively pinpoints particular course steps where learner disengagement is concentrated. Steps that regularly exhibit considerable drop-off over several runs are highlighted in the research, suggesting that disengagement may be a result of structural elements of the course design rather than being solely cohortspecific. Reliable comparisons of engagement patterns between runs with varying cohort sizes are made possible by the use of proportional measures. The analysis's dependence on aggregated interaction data places limitations on it. The underlying causes of dropout cannot be deduced, even though disengagement points are plainly established. External limitations, learner motivation, and topic complexity are examples of factors that are not clearly apparent. Notwithstanding these drawbacks, the findings satisfy the specified success criteria and offer a solid foundation for more research. ### Deployment Targeted assessment and redesign of course steps with consistently high drop-off rates can be guided by the findings. Priority points for enhancing instructional support, pace, or clarity are represented by these phases.The placement of learner support interventions, like reminders or additional resources, at crucial points in the course may also be influenced by the findings. The found patterns can be used as benchmarks in subsequent iterations to assess the efficacy of any modifications made. This facilitates a CRISP-DM framework-aligned iterative improvement approach.

Evaluation

Cycle 1 effectively pinpoints specific course steps where learner disengagement is concentrated. Steps that consistently exhibit substantial drop-off across multiple course runs are highlighted, suggesting that disengagement may stem from structural elements of the course design rather than being solely cohort-specific. The use of proportional engagement measures enables reliable comparisons between runs with varying cohort sizes.

However, the analysis is subject to limitations arising from its reliance on aggregated interaction data. Although disengagement points are clearly identified, the underlying causes of learner dropout cannot be

directly inferred. Factors such as external constraints, learner motivation, and topic complexity are not explicitly captured within the available data. Despite these limitations, the findings meet the predefined success criteria and provide a robust foundation for further investigation.

Deployment

The results from Cycle 1 can guide targeted assessment and redesign of course steps that consistently demonstrate high drop-off rates. These steps represent priority areas for improving instructional clarity, pacing, or learner support.

Additionally, the placement of learner support interventions—such as reminders, prompts, or supplementary learning resources—may be informed by the identified disengagement points. The observed patterns can also serve as benchmarks in future iterations of the course, enabling evaluation of the effectiveness of implemented changes. This supports an iterative, data-driven improvement approach aligned with the CRISP-DM framework.

CRISP-DM Cycle 2 — Assessment Engagement & Persistence

Business Understanding

The second CRISP-DM cycle builds upon the findings of Cycle 1 by examining the relationship between learner persistence and assessment engagement. While Cycle 1 identifies where disengagement occurs, it does not provide early indicators of learners who may be at risk of dropping out.

One plausible behavioural indicator of persistence is active participation in quizzes. Learners who engage with assessments may demonstrate higher commitment to progressing through the course.

The primary research question for Cycle 2 is therefore:

Does participation in quizzes increase learner persistence?

Success in this cycle is defined by identifying whether a statistically significant association exists between quiz engagement and learner persistence, while acknowledging the inherent limitations of observational data.

Data Preparation

Learner-level datasets from all seven course runs are combined to investigate the relationship between persistence and quiz engagement. Previously constructed learner summary tables are used to determine learner persistence, while quiz engagement is quantified using the total number of quiz items attempted by each learner.

To ensure learners who did not participate in quizzes are retained in the analysis, quiz response data are aggregated at the learner level and merged with persistence measures using a left join. Missing quiz counts are replaced with zero to represent non-participation. The resulting analysis-ready dataset includes learner persistence measures, quiz engagement, and run identifiers.

```
learner_quiz_participation <- dplyr::bind_rows(lapply(1:7, function(r) {
lsumm <- get(paste0("learner_summary_run", r)) |>
dplyr::mutate(run = r)
qr <- get(paste0("clean_question_response_run", r)) |>
dplyr::filter(!is.na(correct)) |>
dplyr::group_by(learner_id) |>
dplyr::summarise(
```

```
quiz_items_answered = dplyr::n(),
.groups = "drop"
)
dplyr::left_join(lsumm, qr, by = "learner_id")
})) |>
dplyr::mutate(quiz_items_answered = dplyr::coalesce(quiz_items_answered, 0L))
```

Modelling

To determine if quiz engagement and learner persistence are related, a straightforward linear regression model is employed. The number of quiz items tried is used as the predictor and the completion ratio as the outcome variable. This modeling technique is appropriate for exploratory research because it is purposefully economical and puts interpretability over complexity. The computed slope coefficient shows whether increased course completion behavior is linked to higher quiz participation levels.

```
m1 <- lm(completion_ratio ~ quiz_items_answered,
data = learner_quiz_participation)
knitr::kable(
summary(m1)$coef,
caption = "Linear regression predicting persistence from quiz participation."
)
```

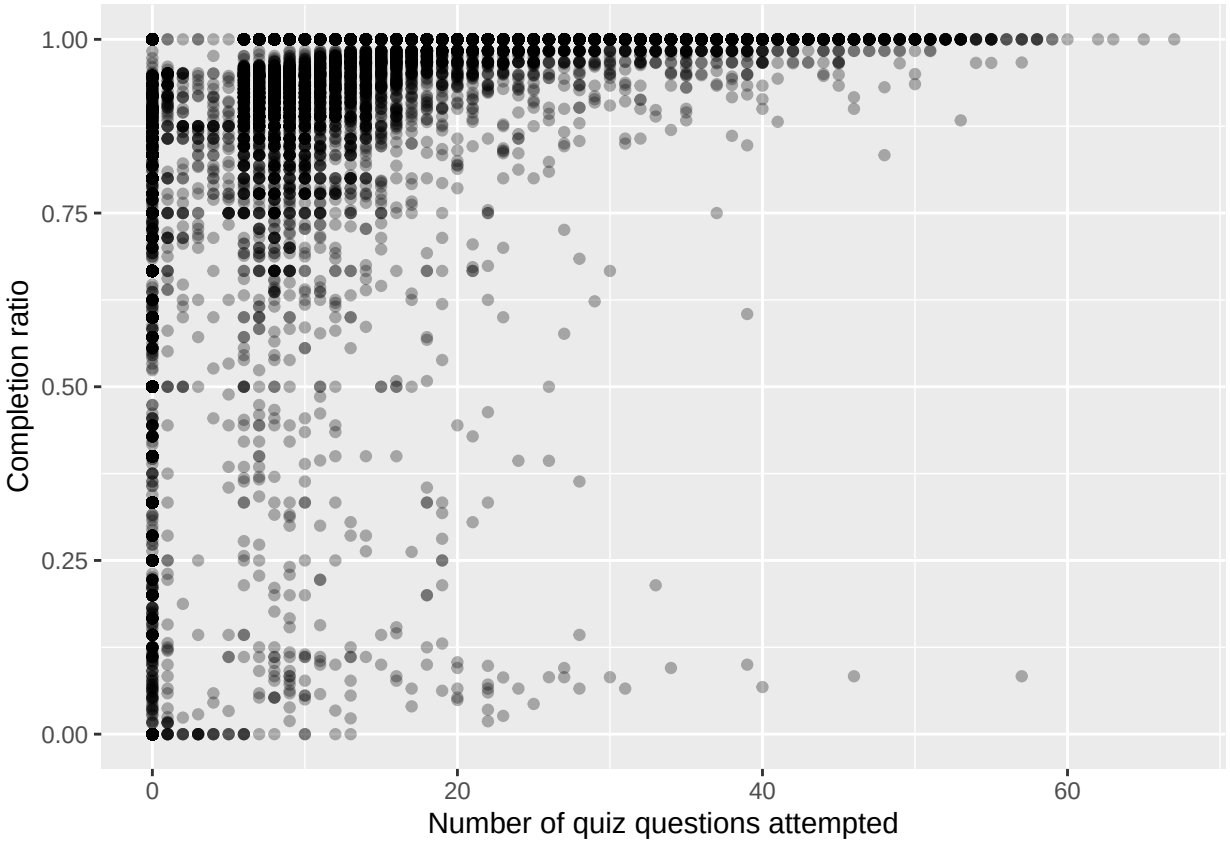
Table 2: Linear regression predicting persistence from quiz participation.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.4302061	0.0028780	149.4805	0
quiz_items_answered	0.0204400	0.0001952	104.7107	0

Correlation Analysis

In addition to regression modelling, the Pearson correlation coefficient is calculated to quantify the strength and direction of the relationship between quiz participation and persistence. This provides a complementary, model-free summary of association and allows for straightforward interpretation of the overall relationship between assessment engagement and learner completion behaviour. The correlation between quiz participation and persistence is r round(cor(learner_quiz_participationquizitemsanswered, learnerquizparticipationcompletion_ratio, use = “complete.obs”), 3).

```
ggplot2::ggplot(
learner_quiz_participation,
ggplot2::aes(
x = quiz_items_answered,
y = completion_ratio
)
) +
ggplot2::geom_point(alpha = 0.3) +
ggplot2::labs(
x = "Number of quiz questions attempted",
y = "Completion ratio"
)
```



Evaluation Cycle 2 shows that learner perseverance and quiz engagement are positively correlated. The use of quiz participation as a potential early predictor of persistent engagement is supported by the fact that students who participate in assessments are more likely to continue moving through the course. By going beyond detecting disengagement points to discovering behavioral patterns linked to persistence, this enhances the results from Cycle 1. Care should be taken while interpreting the results. Because the research is observational, it is vulnerable to survivorship bias because students who continue to participate are more likely to take tests. Furthermore, unseen elements like learner motivation, past knowledge, or outside obligations may have an impact on perseverance and assessment participation. Notwithstanding these drawbacks, the outcomes meet Cycle 2's success criteria and offer insightful information about learner behavior.

Deployment

Practical initiatives targeted at enhancing student retention can be informed by the findings of Cycle 2. To identify students who are at danger of disengagement, early-warning systems may use quiz engagement data. Reminders or more learning materials could then be used to target students who are participating in assessments at a low or declining rate. These insights complement a data-driven approach to course improvement when paired with the step-level findings from Cycle 1. While assessment engagement can be used to track learner persistence throughout subsequent course runs, structural improvements can be prioritized at identified disengagement points.

Limitations

Several limitations should be acknowledged when interpreting the findings of this study. First, the analysis relies on aggregated interaction data, which restricts the ability to infer causal relationships between learner

behaviour and disengagement. While step-level drop-off points and associations with quiz engagement are clearly identified, the underlying reasons for these behaviours cannot be directly observed.

Second, the study is based on observational data and is therefore subject to survivorship bias. Learners who remain active in the course are more likely to participate in quizzes, which may inflate the observed association between assessment engagement and persistence. Additionally, important learner-specific factors such as motivation, prior subject knowledge, learning goals, and external commitments are not captured in the available datasets.

Finally, the analysis focuses on behavioural indicators within the platform and does not incorporate qualitative feedback or contextual information about course content, instructional design, or learner experience. As a result, some disengagement patterns may be influenced by factors that are not measurable within the current analytical framework.

Future Work

Future work could extend this analysis in several directions. Incorporating richer learner-level features, such as temporal engagement patterns or time spent on individual steps, may provide deeper insight into disengagement mechanisms. The use of sequence-based or survival analysis methods could also improve understanding of when learners are most at risk of dropping out.

In addition, future studies could integrate qualitative data, such as learner surveys or feedback, to better contextualise the quantitative findings and identify specific instructional challenges. Experimental or quasi-experimental designs, such as A/B testing of redesigned course steps, could help establish causal links between course modifications and improved learner retention.

Finally, the predictive potential of quiz engagement could be further explored by developing early-warning models that combine assessment participation with step-level engagement metrics. Such models could support real-time interventions aimed at improving learner persistence across future course runs.

Final Conclusions

This study finds step-level points of learner disengagement over two CRISP-DM cycles and demonstrates that quiz engagement is positively correlated with learner perseverance. The findings offer evidence-based recommendations for enhancing learner assistance techniques and course design. While assessment engagement is a useful signal for identifying learners at risk of disengagement, step-level analysis identifies areas where structural changes may be necessary. All things considered, the analysis highlights the importance of a methodical, data-driven approach to comprehending and enhancing participation in extensive online courses