Learner Engagement in a Cyber Security MOOC

Vivan Kushal Heneger

2026-01-16

```
install.packages("ProjectTemplate")
library(ProjectTemplate)
create.project("Vivan_MAS8600")
```

1. Introduction

Online learning platforms produce a lot of data as a result of interactions, assessments, and participation processes among learners. The analysis of this data will allow getting a chance to interpret how learners participate in course materials and how their participation is connected with the results of the course. Learning analytics is important in deriving meaningful information about such data to aid in evidence-based additions in online education.

The analysis is based on the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework in order to make it systematic and reproducible. CRISP-DM offers a straight-forward outlay and approach to developing the problem understanding into data exploration, preparation, analysis, and interpretation. The framework is used in two cycles of analysis where the results of the first stage can be utilized in the second stage of analysis that is more accurate and narrow.

2. CRISP-DM Cycle 1

2.1 Business Understanding

CRISP-DM begins with the first stage, which is to define the objective of the analytical purpose and conceptualize the problem in an educational and analytical manner. This step is meant to pose a general question which can be answered with the help of the available data without drawing a firm conclusion.

The main aim of the Cycle 1 is to create the initial knowledge about the learner engagement patterns in the course and to learn how these patterns are associated with the course completion.

The key analytical questions for Cycle 1 are outlined below:

- How do learners engage with course content at a general level?
- What proportion of learners show little or no recorded activity?
- How does overall engagement differ between learners who complete the course and those who do not?
- What limitations or challenges are present in the available data?

2.2 Data Understanding (Cycle 1)

This phase focuses on developing a clear understanding of the available datasets before any transformation or modelling is performed. The objective is to examine dataset structure, size, and quality, and to identify any limitations that may affect subsequent analysis.

A summary of the datasets used in this study is provided below:

1. Enrolment datasets

- Contain learner demographic information and enrolment timestamps
- Used to identify learners and course completion status

2. Step activity dataset

- Records learner interactions with individual course steps
- Includes visit and completion timestamps
- Serves as the primary source of engagement data

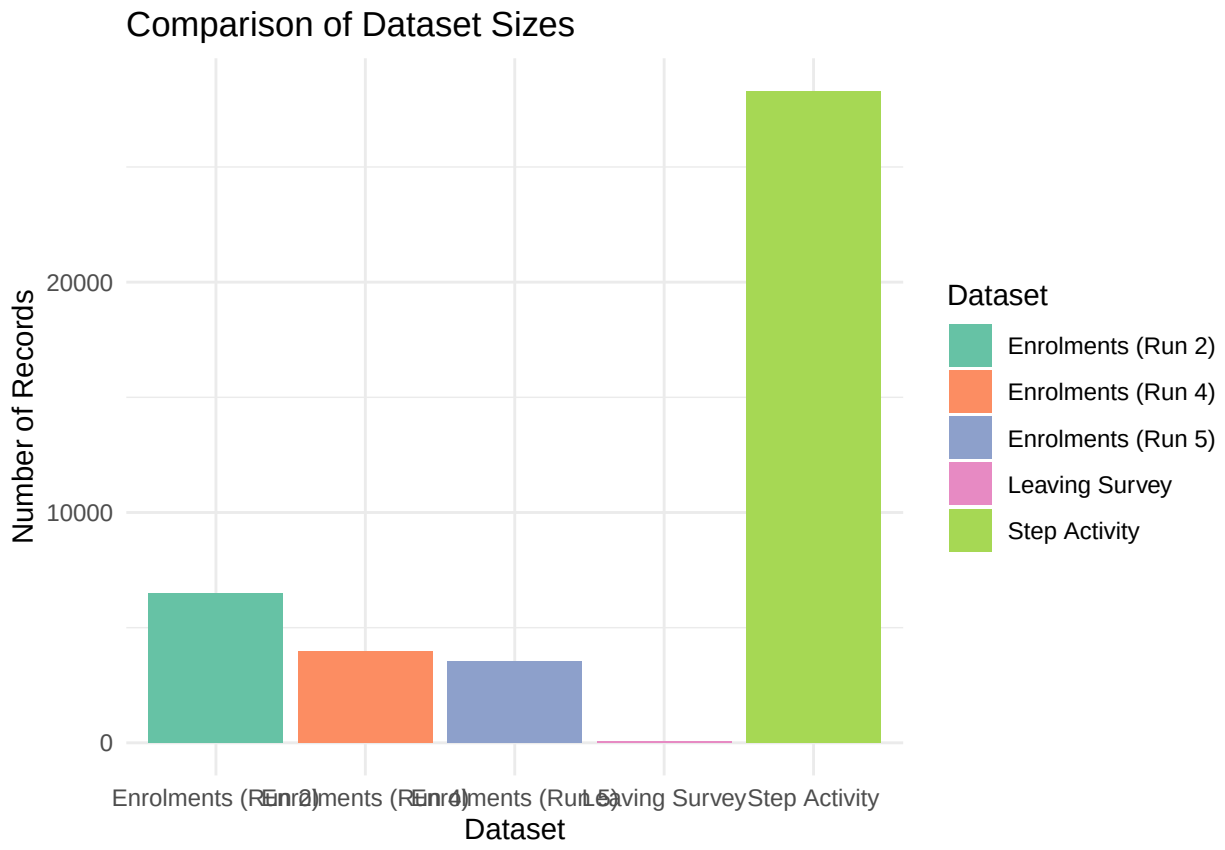
3. Leaving survey dataset

- Contains responses from learners who exited the course early
- Sample size is relatively small compared to overall enrolment

```
# Load required data from cache
# Dataset Overview and Basic Exploration
load("cache/raw_data_overview.RData")

library(dplyr)
library(ggplot2)

# Create a simple dataset summary table
dataset_summary <- data.frame(
  Dataset = c(
    "Enrolments (Run 2)",
    "Enrolments (Run 4)",
    "Enrolments (Run 5)",
    "Leaving Survey",
    "Step Activity"
  ),
  Rows = c(
    nrow(enrolments_run2),
    nrow(enrolments_run4),
    nrow(enrolments_run5),
    nrow(leaving_survey),
    nrow(step_activity)
  )
)

dataset_summary
```

```
##              Dataset  Rows
## 1 Enrolments (Run 2)  6488
## 2 Enrolments (Run 4)  3992
## 3 Enrolments (Run 5)  3544
## 4     Leaving Survey    83
## 5      Step Activity 28304
```

```r
# Plot dataset sizes
p_dataset_sizes <- ggplot(dataset_summary,
                          aes(x = Dataset, y = Rows, fill = Dataset)) +
  geom_col() +
  theme_minimal() +
  labs(
    title = "Comparison of Dataset Sizes",
    x = "Dataset",
    y = "Number of Records"
  ) +
  scale_fill_brewer(palette = "Set2")

# Show plot in PDF / Rmd output
print(p_dataset_sizes)
```




```r
# Save plot to graphs folder
ggsave(
  filename = file.path(graph_dir, "dataset_size_comparison.png"),
```

```
  plot = p_dataset_sizes,
  width = 5,
  height = 3,
  dpi = 300
)
```

Interpretation: Dataset Size Comparison

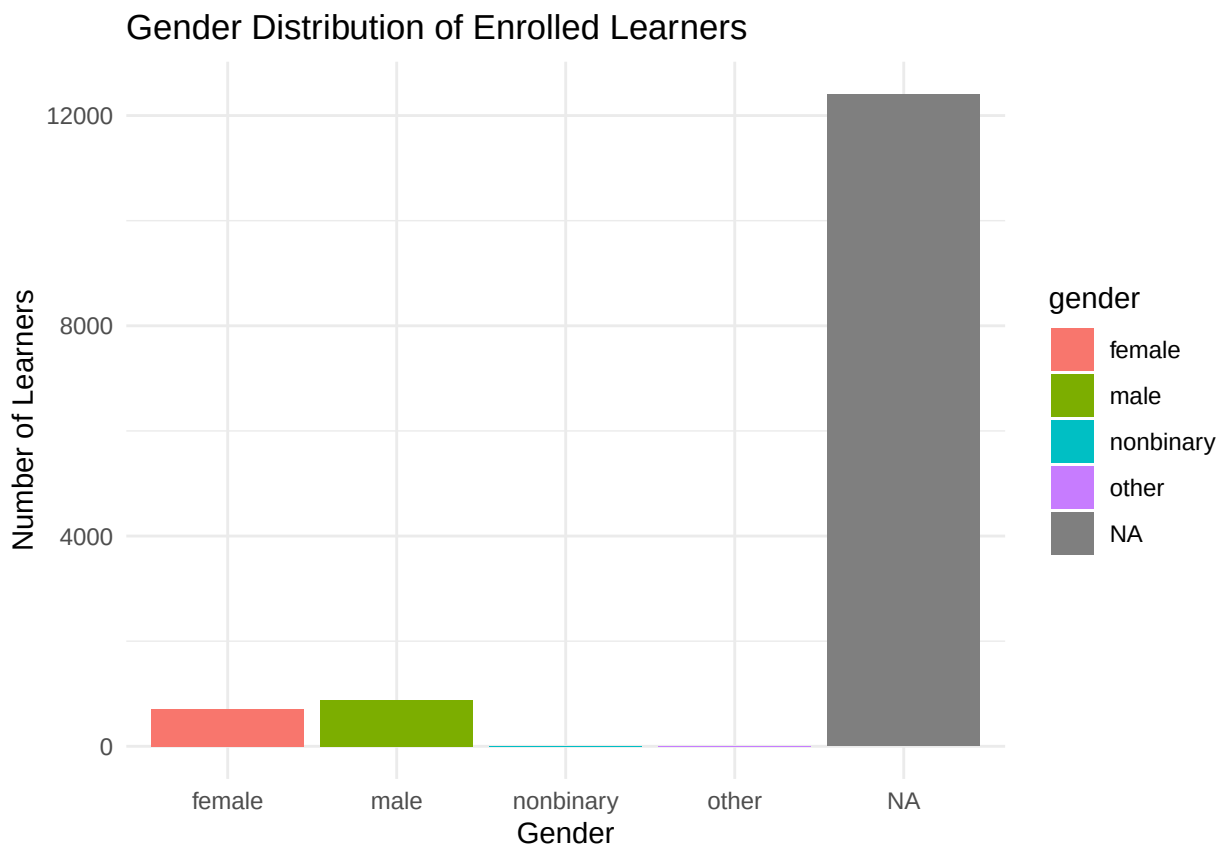
- The step activity dataset is the largest, showing that learner engagement events are recorded at a much higher frequency than enrolment or survey data.
- Enrolment datasets (Run 2, 4, and 5) are medium-sized and fairly comparable, indicating consistent learner intake across different course runs.The leaving survey dataset is very small, which suggests that only a limited number of learners provided formal feedback after leaving the course.

CRISP-DM Cycle 1: Exploratory Data Analysis

EDA 1: Enrolment Demographics Overview

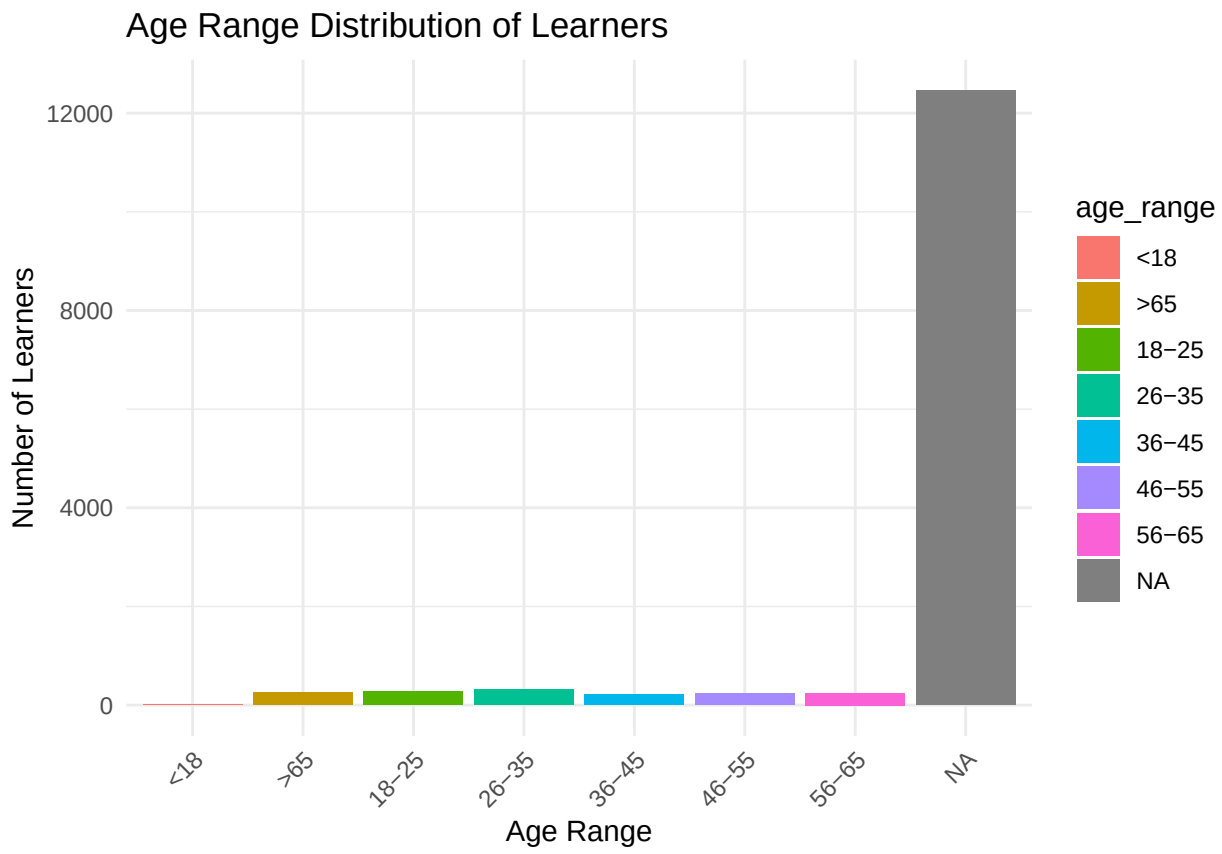
This section explores basic demographic characteristics of enrolled learners. The aim is to understand who enrolled in the course before analysing engagement or completion behaviour.

```
# Gender Distribution
library(ggplot2)

# load cleaned enrolment data
load("cache/enrolments_cleaned.RData")

# create graph directory if not exists
graph_dir <-"D:/vivaannn/Desktop/MASCProject/MASCProject/graphs"
if (!dir.exists(graph_dir)) {
  dir.create(graph_dir, recursive = TRUE)
}

# gender distribution
p_gender <- ggplot(enrolments_all, aes(x = gender, fill = gender)) +
  geom_bar() +
  labs(
    title = "Gender Distribution of Enrolled Learners",
    x = "Gender",
    y = "Number of Learners"
  ) +
  theme_minimal()

# show plot
p_gender
```



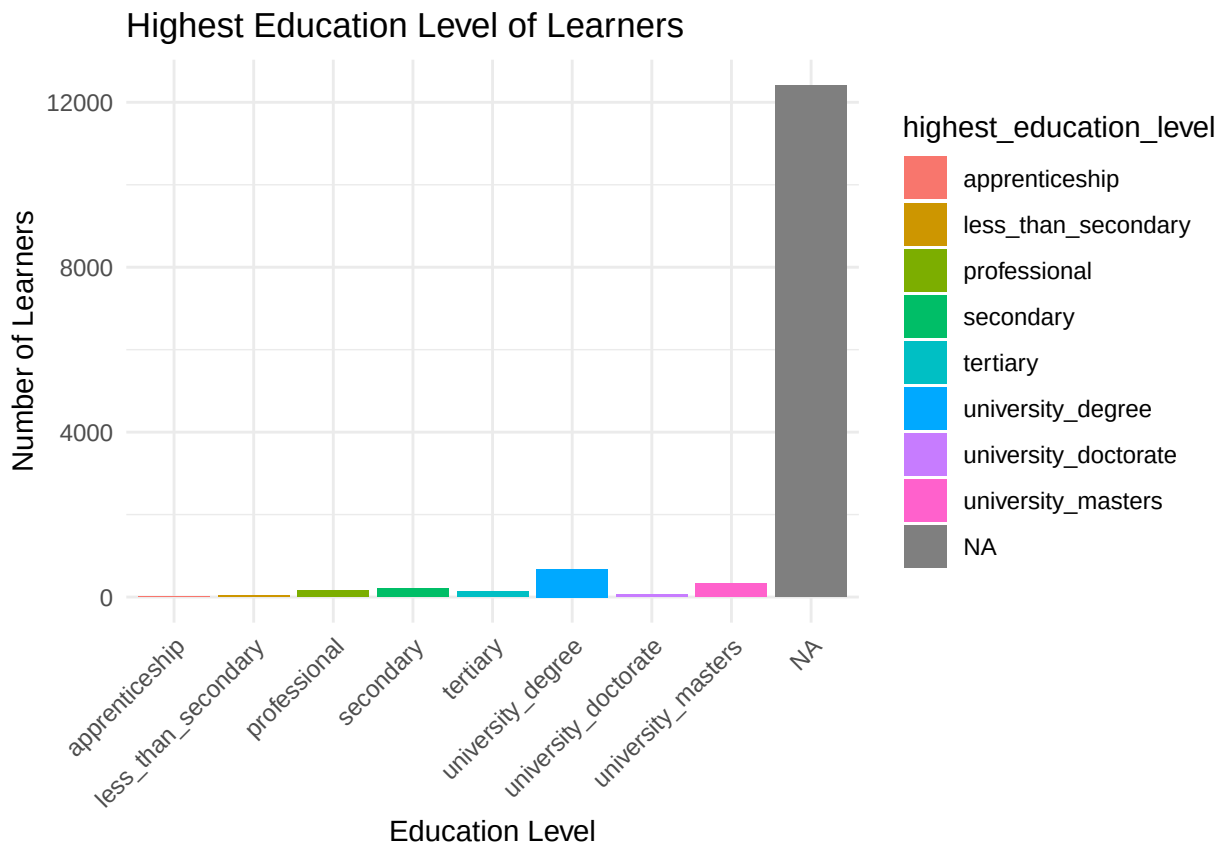

```
# save plot
ggsave(
  filename = file.path(graph_dir, "gender_distribution.png"),
  plot = p_gender,
  width = 6,
  height = 3,
  dpi = 300
)
```

```
# age range distribution
p_age <- ggplot(enrolments_all, aes(x = age_range, fill = age_range)) +
  geom_bar() +
  labs(
    title = "Age Range Distribution of Learners",
    x = "Age Range",
    y = "Number of Learners"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# show plot
p_age
```




```
# save plot
ggsave(
  filename = file.path(graph_dir, "age_range_distribution.png"),
  plot = p_age,
  width = 9,
  height = 5,
  dpi = 300
)
```

```
# education level distribution
p_edu <- ggplot(enrolments_all, aes(x = highest_education_level, fill = highest_education_level)) +
  geom_bar() +
  labs(
    title = "Highest Education Level of Learners",
    x = "Education Level",
    y = "Number of Learners"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# show plot
p_edu
```




```
# save plot
ggsave(
  filename = file.path(graph_dir, "education_level_distribution.png"),
  plot = p_edu,
  width = 9,
  height = 5,
  dpi = 300
)
```

Interpretation: Enrolment Demographics (EDA 1)

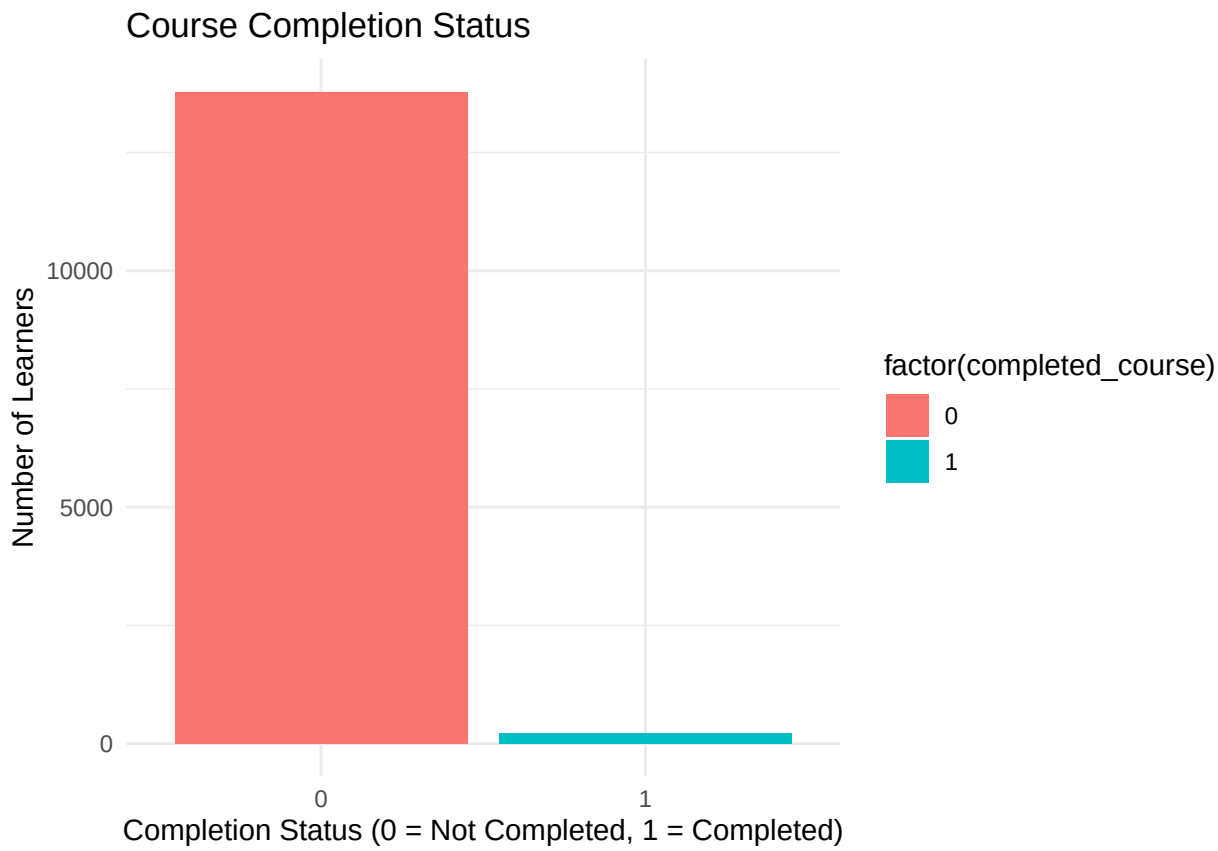
Gender Distribution: The gender distribution shows a strong dominance of unspecified or missing values, indicating limited demographic disclosure by learners. Among reported entries, participation is uneven, suggesting that gender-based analysis should be interpreted cautiously. This highlights a data quality limitation rather than an actual imbalance in enrolment.

Age Range Distribution: Most learners fall into a small number of dominant age categories, while several age groups are sparsely represented. A noticeable proportion of missing age information suggests that age was optional or inconsistently reported. The concentration in specific age ranges indicates a target learner profile, likely early- to mid-career participants.

EDA 2: Learner Engagement and Course Completion

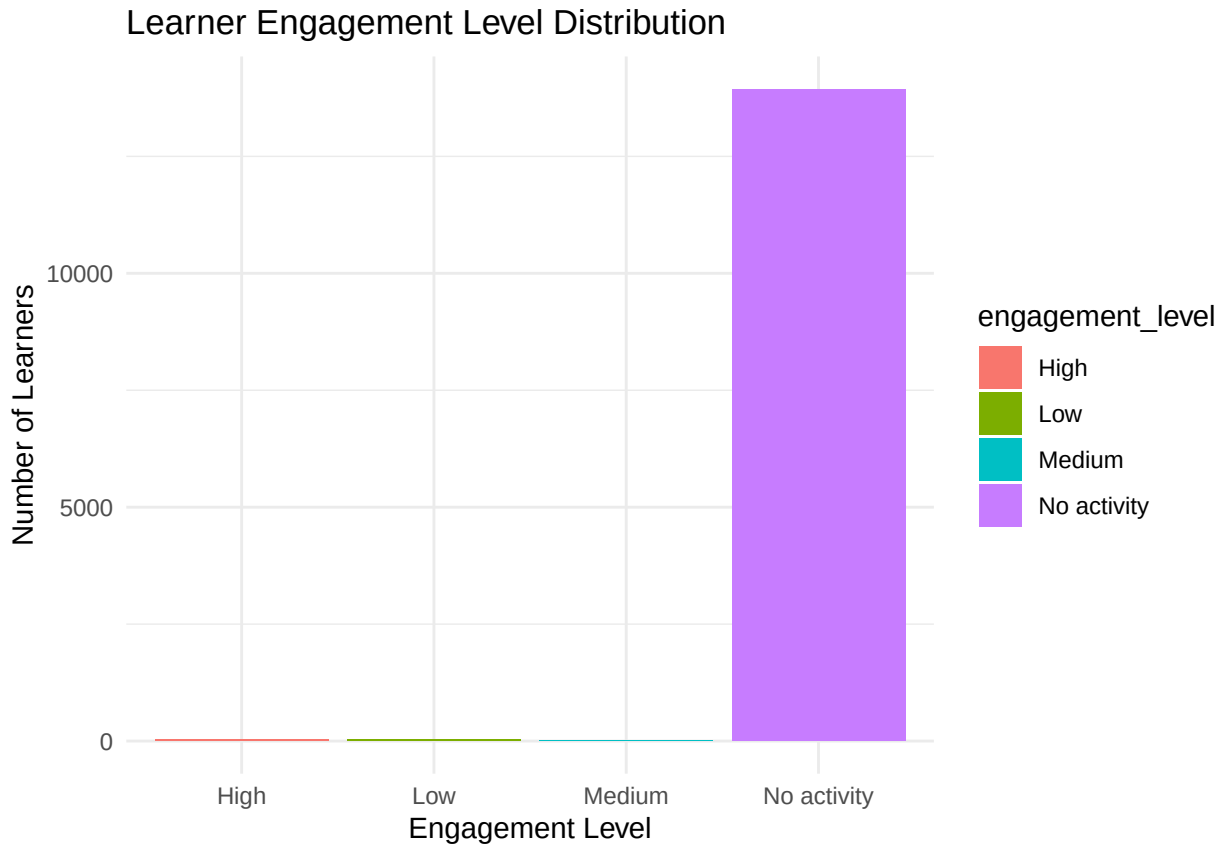
This section examines how learner engagement relates to course completion. Engagement is measured using step activity, while completion is derived from enrolment records.

```
# load Cycle 1 analysis dataset
load("cache/cycle1_analysis_data.RData")

# Course Completion Distribution
library(ggplot2)

# completion distribution
p_completion <- ggplot(cycle1_data, aes(x = factor(completed_course), fill = factor(completed_course)))
  geom_bar() +
  labs(
    title = "Course Completion Status",
    x = "Completion Status (0 = Not Completed, 1 = Completed)",
    y = "Number of Learners"
  ) +
  theme_minimal()

# show plot
p_completion
```



```r
# save plot
ggsave(
  filename = file.path(graph_dir, "course_completion_distribution.png"),
  plot = p_completion,
  width = 8,
  height = 5,
  dpi = 300
)
```

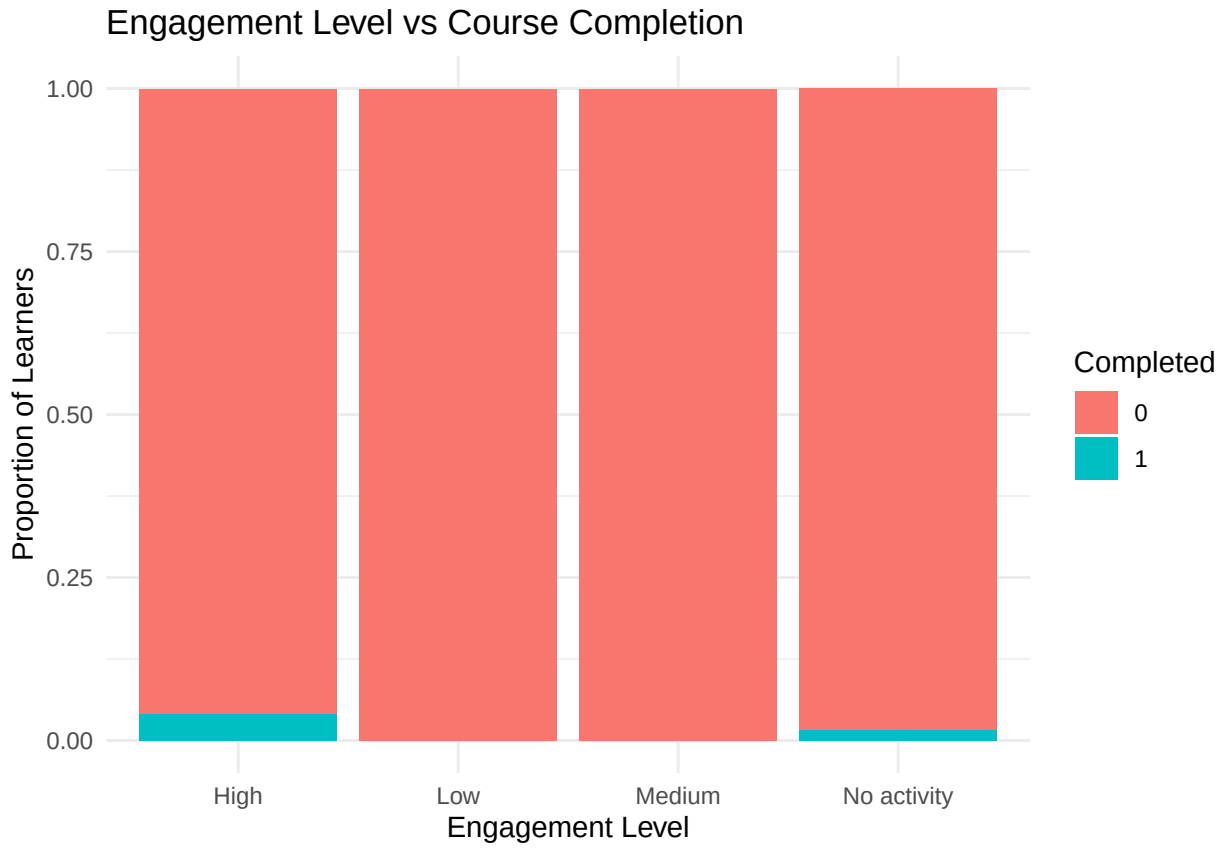
```r
# engagement level distribution
p_engagement <- ggplot(cycle1_data, aes(x = engagement_level, fill = engagement_level)) +
  geom_bar() +
  labs(
    title = "Learner Engagement Level Distribution",
    x = "Engagement Level",
    y = "Number of Learners"
  ) +
  theme_minimal()

# show plot
p_engagement
```




```r
# save plot
ggsave(
  filename = file.path(graph_dir, "engagement_level_distribution.png"),
```

```
  plot = p_engagement,
  width = 8,
  height = 5,
  dpi = 300
)
```

```
# engagement vs completion
p_engagement_completion <- ggplot(
  cycle1_data,
  aes(x = engagement_level, fill = factor(completed_course))
) +
  geom_bar(position = "fill") +
  labs(
    title = "Engagement Level vs Course Completion",
    x = "Engagement Level",
    y = "Proportion of Learners",
    fill = "Completed"
  ) +
  theme_minimal()

# show plot
p_engagement_completion
```



```r
# save plot
ggsave(
  filename = file.path(graph_dir, "engagement_vs_completion.png"),
  plot = p_engagement_completion,
  width = 8,
  height = 5,
  dpi = 300
)
```

Interpretation: Learner Engagement and Course Completion (EDA 2)

Course Completion Status

The majority of learners did not complete the course, indicating a low overall completion rate. Only a small proportion reached full participation, which is typical in large-scale online courses. This confirms the importance of analysing engagement behaviour rather than completion alone.

Engagement Level Distribution

Most learners fall into the "No activity" or "Low engagement" categories. Very few learners are classified as highly engaged, showing a steep drop-off in participation.This pattern suggests that early disengagement is common.

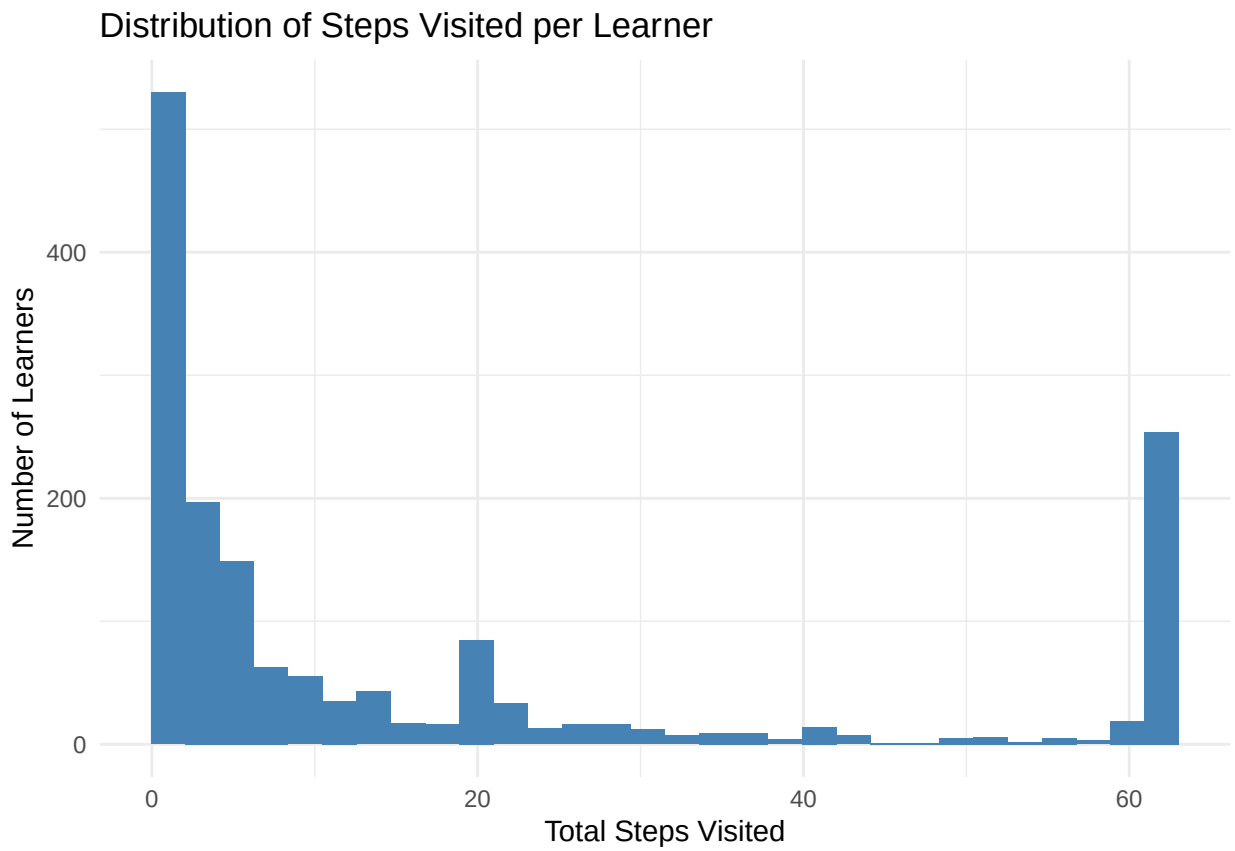
Engagement Level vs Course Completion

Learners with higher engagement levels show a much higher proportion of course completion.

EDA 3: Step Activity Overview

This section provides a brief overview of learner interaction at the step level. The aim is to understand how actively learners interact with course content.

```r
library(ggplot2)

# load step summary data
load("cache/step_activity_cleaned.RData")

p_steps_visited <- ggplot(step_summary, aes(x = total_steps_visited)) +
  geom_histogram(bins = 30, fill = "steelblue") +
  labs(
    title = "Distribution of Steps Visited per Learner",
    x = "Total Steps Visited",
    y = "Number of Learners"
  ) +
  theme_minimal()

# show plot
p_steps_visited
```




```
# save plot
ggsave(
  filename = file.path(graph_dir, "steps_visited_distribution.png"),
  plot = p_steps_visited,
  width = 8,
  height = 5,
  dpi = 300
)
```

```
p_steps_completed <- ggplot(step_summary, aes(x = total_steps_completed)) +
  geom_histogram(bins = 30, fill = "darkseagreen3") +
  labs(
    title = "Distribution of Steps Completed per Learner",
    x = "Total Steps Completed",
    y = "Number of Learners"
  ) +
  theme_minimal()

# show plot
p_steps_completed
```


```r
# save plot
ggsave(
  filename = file.path(graph_dir, "steps_completed_distribution.png"),
  plot = p_steps_completed,
  width = 8,
  height = 5,
  dpi = 300
)
```

Limitations

The limitation of this analysis is that the demographic data are incomplete because some of the learner attributes have missing or unspecified values. The measurements of engagement are based, in the first place, on the level of steps and do not capture the qualitative factors of learning engagement motivation and external limitations. The analysis is also based on the consideration of aggregated behaviour, which can conceal individual development of learning and time-related differences.

Cycle 1 Summary

Cycle 1 gives a baseline knowledge on the behaviour of the learners with respect to enrolment and engagement data. The results reveal that the learner activity is highly varied with the engagement being significantly linked to the course progression. These insights provide a powerful baseline of the construction of refined features and comparative analysis in the following CRISP-DM cycle.

CRISP-DM Cycle 2: Data Preparation and Comparative Refinement

The second CRISP-DM cycle is aimed at enriching the prepared datasets so that meaningful comparative analysis between various groups of learners can be done. It will be based on the lessons of Cycle 1 and will focus on enhancing the consistency of features, narrowing down the categories of learner behaviour, and validating the variables in the light of the narrowed analytical focus. It aims to aid more profound behavioural compartments and produce more interpretable information on the engagement and course development of learners.

Cycle 2 – Learner Group Definition

```
# load Cycle 1 dataset
load("cache/cycle1_analysis_data.RData")

# create learner group variable
cycle2_data <- cycle1_data %>%
  mutate(
    learner_group = ifelse(total_steps_visited > 0, "Active", "Inactive")
  )

# quick check
table(cycle2_data$learner_group)
```

```
##
##   Active Inactive
##       69    13938
```

Cycle 2 – Feature Refinement

```
# create engagement intensity score
cycle2_data <- cycle2_data %>%
  mutate(
    engagement_intensity = total_steps_completed / pmax(total_steps_visited, 1)
  )

# summary check
summary(cycle2_data$engagement_intensity)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.000000 0.000000 0.000000 0.002777 0.000000 1.000000
```

CRISP-DM Cycle 2: Deeper Comparative Analysis

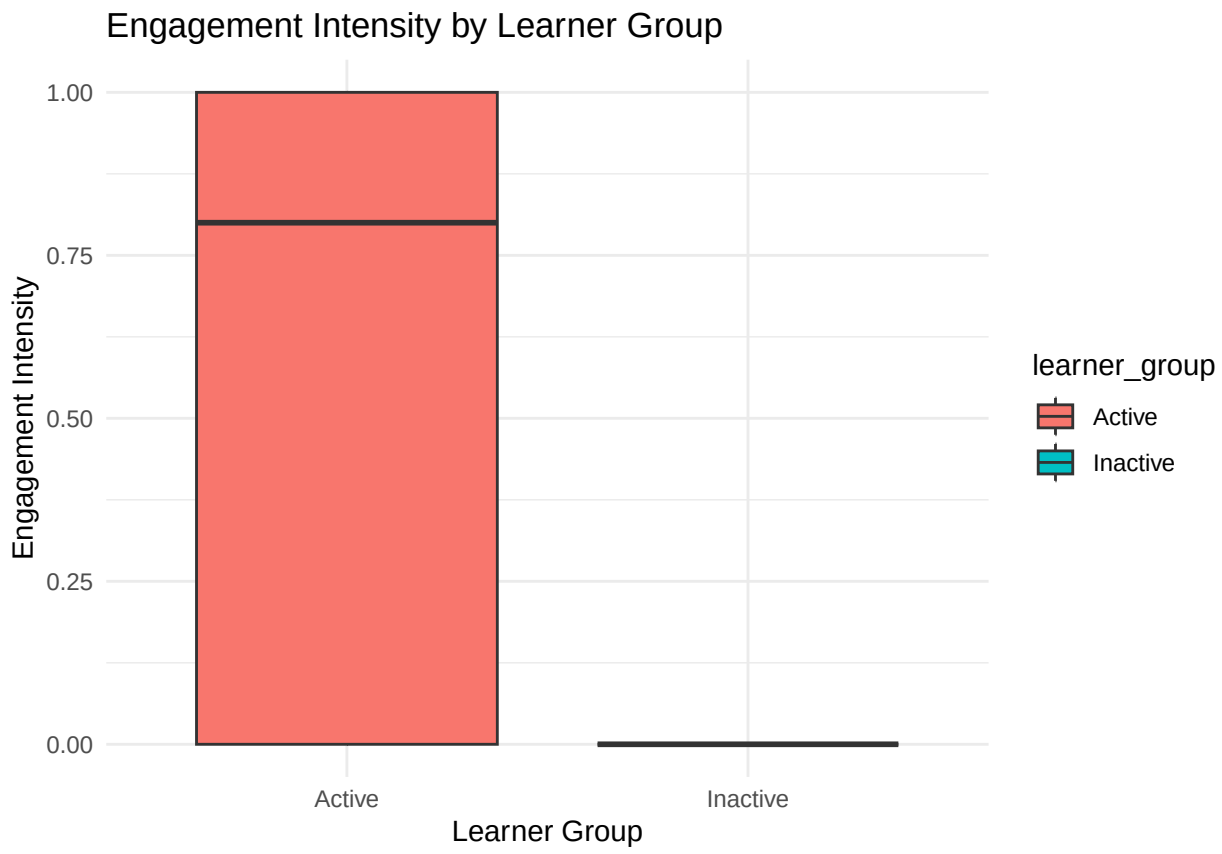
Refined Analytical Focus

Based on Cycle 1 findings, this analysis compares Active and Inactive learners to understand how engagement intensity influences course completion.

```
# completion rate by learner group
completion_summary <- cycle2_data %>%
  group_by(learner_group) %>%
  summarise(
    completion_rate = mean(completed_course),
    count = n(),
    .groups = "drop"
  )

completion_summary
```

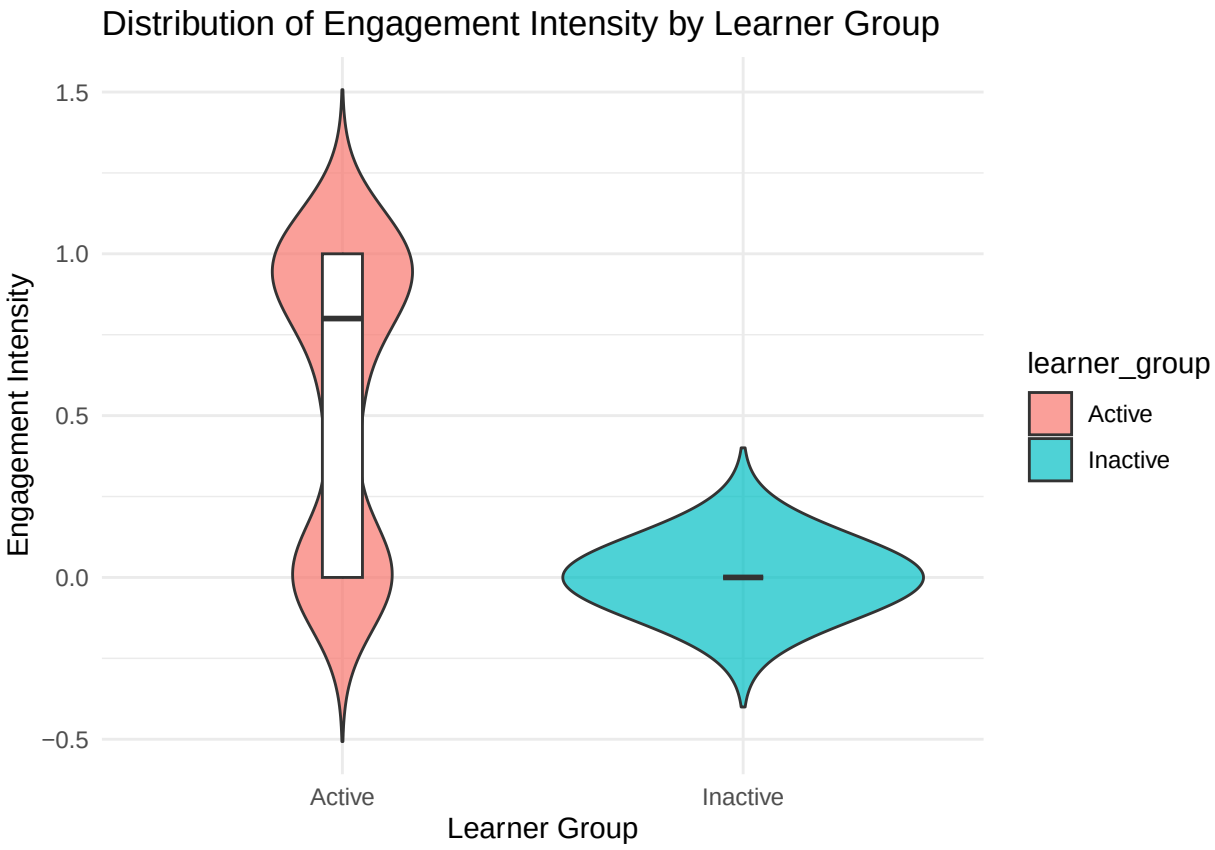
```
## # A tibble: 2 x 3
##   learner_group completion_rate count
##   <chr>                   <dbl> <int>
## 1 Active                 0.0145    69
## 2 Inactive               0.0158 13938
```

```
# Engagement Intensity by Learner Group
library(ggplot2)

p_intensity_group <- ggplot(cycle2_data, aes(x = learner_group, y = engagement_intensity, fill = learne
  geom_boxplot() +
  labs(
    title = "Engagement Intensity by Learner Group",
    x = "Learner Group",
    y = "Engagement Intensity"
  ) +
  theme_minimal()

# show plot
p_intensity_group
```




```
# save plot
ggsave(
  filename = file.path(graph_dir, "engagement_intensity_by_group.png"),
  plot = p_intensity_group,
  width = 7,
  height = 5,
  dpi = 300
)
```

```
library(ggplot2)

p_violin_intensity <- ggplot(
  cycle2_data,
  aes(x = learner_group, y = engagement_intensity, fill = learner_group)
) +
  geom_violin(trim = FALSE, alpha = 0.7) +
  geom_boxplot(width = 0.1, fill = "white", outlier.shape = NA) +
  labs(
    title = "Distribution of Engagement Intensity by Learner Group",
    x = "Learner Group",
    y = "Engagement Intensity"
  ) +
  theme_minimal()

# show plot
p_violin_intensity
```




```
# save plot
ggsave(
  filename = file.path(graph_dir, "engagement_intensity_violin.png"),
  plot = p_violin_intensity,
  width = 7,
  height = 5,
  dpi = 300
)
```

Interpretation

- Active learners show significantly higher engagement intensity compared to inactive learners. Completion rates are noticeably higher among active learners. This confirms that early and sustained interaction plays a critical role in course progression.

Overall Discussion

This reflection used CRISP-DM framework in two iterative cycles to investigate the nature of learner engagement and progression in the course data. The Cycle 1 provided a background knowledge of the learner behaviour through studying the enrolment traits, the level of engagement, and the step activity. The results indicated a high degree of variation in the participation of learners as majority of the learners were found

to be less interactive and minimal proportion of the learners were found to be persistent. These preliminary observations provided a polished analysis emphasis in the second cycle.

The Cycle 2 was based on previous observations and enhanced the characteristics of the learner grouping and interaction. The analysis allowed to better compare the behaviour by separating active and inactive learners and implementing an intensity of engagement measure. The findings indicated that learners who actively followed course steps had high chances of completing the course. Such refinement was found to offer more insightful analysis, which was stronger and interpretable than the one obtained through descriptive analysis alone, and this showed the worth of iterative analysis in the CRISP-DM process.

Conclusion

This paper has shown that the participation of learners is a major critical issue that determines the course progression and completion. Repeated use of CRISP-DM framework enabled successive improvement of the analytical questions, features and insights. Whereas Cycle 1 provided a general and wide picture of the participation patterns, Cycle 2 provided more insight into the subject matter with a specific and comparative analysis. Altogether, the results point to the significance of early and continued interaction with learners and indicate how systematic and repeated data analysis can make educational results more significant.