Learner Persistence by Country

Krishnanand Sagar

2026-01-13

1. Business Understanding - Cycle 1

The objective of this analysis is to understand how learner persistent varies across countries in an online course from a provider perspective identifying countries where learners remain engaged for longer verses those where learners drop out early can inform strategic decisions across course delivery market focus and engagement interventions

Specifically this cycle aims: Compare learner persistent across Identify countries with high early Provide evidence to support targeted engagement strategies in later analysis cycles

2. Data Understanding

The analysis uses country level persistent metrics derived from learner step activity data each learner's engagement is represented by the maximum step reached during the course capturing how far they progressed before disengaging Key variables used in the Analysis include Country detected learner country Learner's number of learners per country Average max step average maximum step reached by learners In the country Pearly drop out rate proportion of learners disengaging within the first two steps These variables allow both overall persistent and early disengagement behaviour to be examined at the country level

3. Data Preparation

Learner step activity data were cleaned with philtre to remove missing country information for each learner the maximum step reached was calculated to represent persistence these learner level values were then aggregated by contrary to produce summary matrix including average persistence and early dropout rate the resulting country level data set was saved to disc and used consistently across all analysis to ensure reproducibility and separation between data preparation and analysis stage

4. Analysis - Cycle 1

The first plot examines variation in learner persistent across countries by focusing on the 10 countries with the largest number of learners using average maximum step reached this visualisation highlights how far learners typically progress in different region while ensuring comparisons are based on sufficiently large learner populations the second plot focuses on early dropout behaviour showing the proportion of learners who disengage early versus those who remain active beyond the initial steps countries are ordered by early dropout rate allowing direct comparison of engagement composition across regionstogether these plots provide complementary perspective one summarising overall persistence depth and other revealing how engagement breaks down at the earliest stage of the course

```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.5.2
```

```
## Warning: package 'ggplot2' was built under R version 4.5.2
```

```
## Warning: package 'dplyr' was built under R version 4.5.2
```

```
## -- Attaching core tidyverse packages ----------------------- tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.1     v stringr   1.5.2
## v ggplot2   4.0.1     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
country_metrics<-read_csv("../cache/country_metrics_cycle1.csv")
```

```
## Rows: 190 Columns: 5
## -- Column specification ------------------------------------------------------
## Delimiter: ","
## chr (1): country
## dbl (4): learners, avg_max_step, median_max_step, early_dropout_rate
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
country_metrics %>%
  slice_max(learners, n = 10) %>%   # top 10 by volume, not performance
  ggplot(aes(x = reorder(country, avg_max_step), y = avg_max_step)) +
  geom_boxplot() +
  coord_flip() +
  labs(
    title = "Variation in Learner Persistence Across Countries",
    x = "Country",
    y = "Average Maximum Step Reached"
  )
```
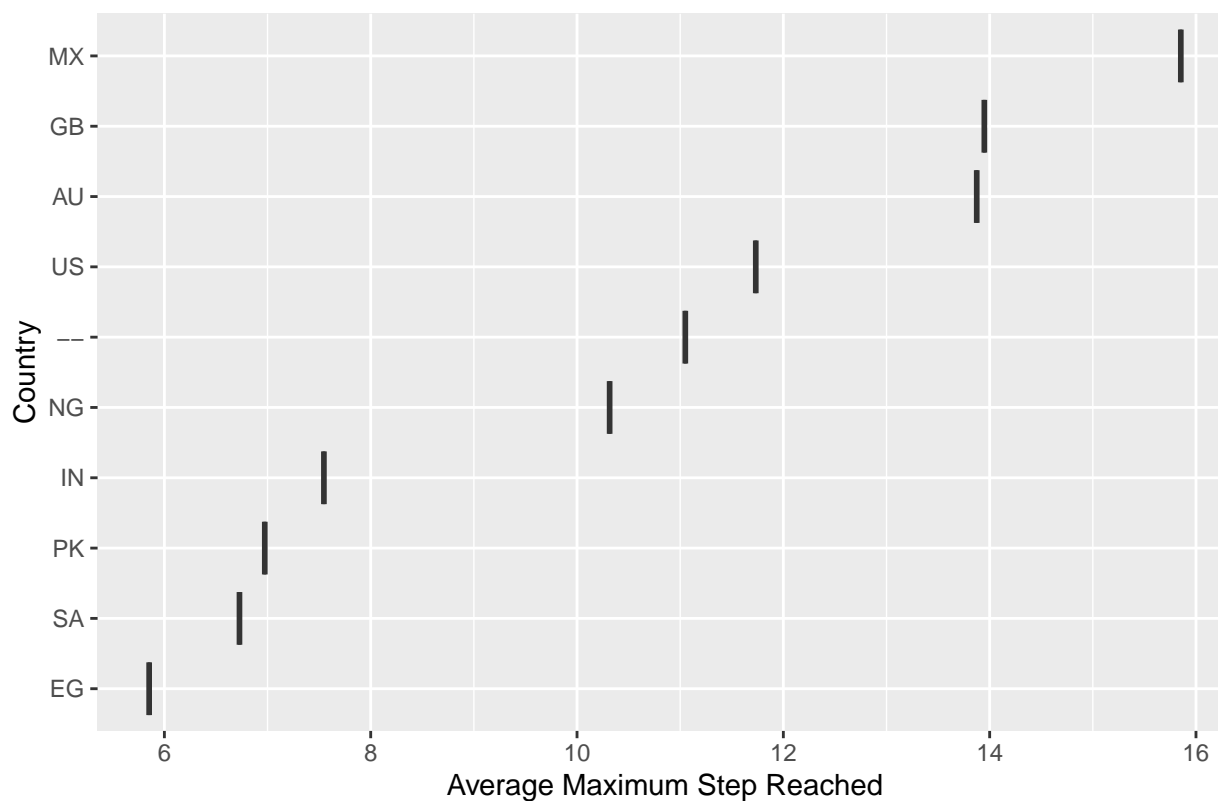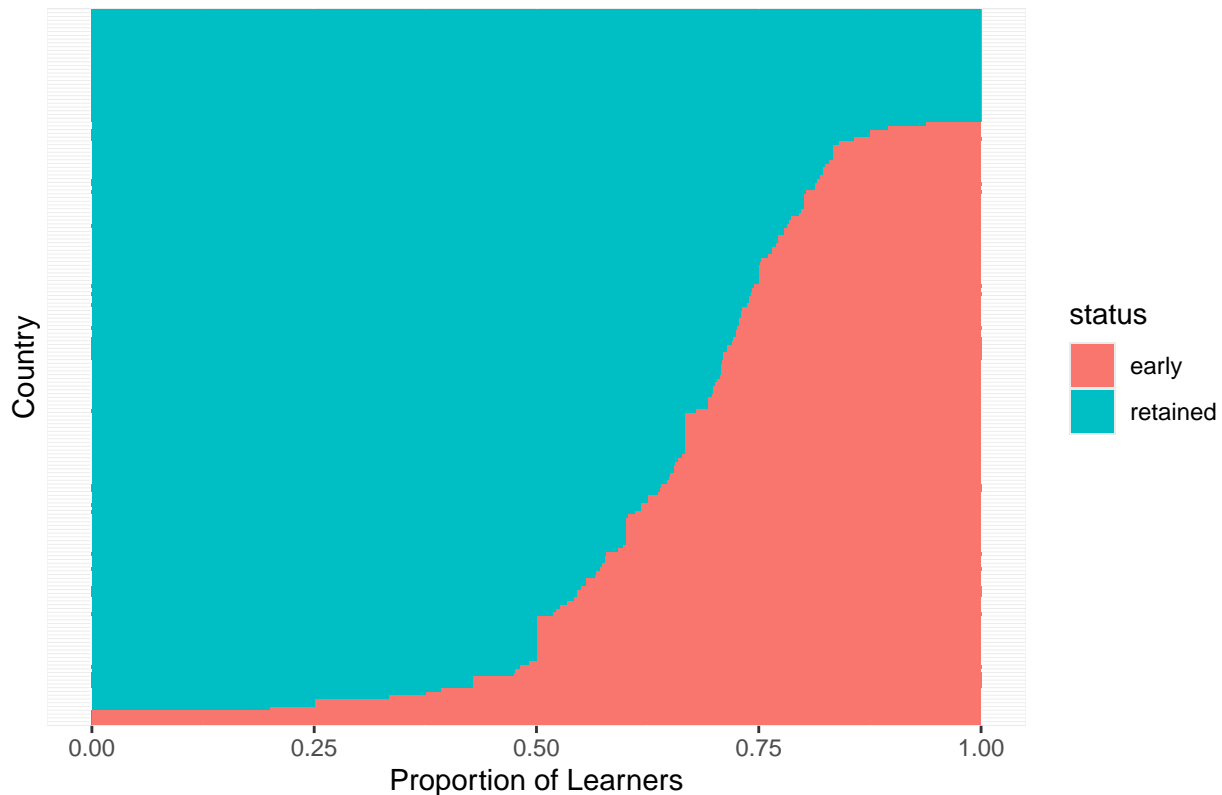
Variation in Learner Persistence Across Countries


```r
country_metrics %>%
  arrange(desc(early_dropout_rate)) %>%
  mutate(
    country = factor(country, levels = country),
    early = early_dropout_rate,
    retained = 1 - early_dropout_rate
  ) %>%
  select(country, early, retained) %>%
  pivot_longer(
    cols = c(early, retained),
    names_to = "status",
    values_to = "proportion"
  ) %>%
  ggplot(aes(
    x = country,
    y = proportion,
    fill = status
  )) +
  geom_col() +
  coord_flip() +
  labs(
    title = "Early Dropout vs Retained Learners by Country",
    x = "Country",
    y = "Proportion of Learners"
  ) +
  theme(
```

```
    axis.text.y = element_blank(),
    axis.ticks.y = element_blank()
)
```

Early Dropout vs Retained Learners by Country

5. Evaluation - Cycle 1

The analysis demonstrates clear difference in learner persistent patterns across country some countries show relatively high average progression through the course indicating sustained learner engagement while other exhibit high early dropout rates suggesting disengagement occurs very quickly From a business perspective Countries with stronger persistent represent stable markets for fully online course delivery Country with high early dropouts may face barriers such as accessibility expectations mismatch or insufficient early support

These findings justify further investigation in cycle two were targeted strategies can be evaluated to improve engagement and reduce early dropout in low persistent regions

Cycle 2: Business Understanding

The objective of Cycle 2 is to translate the country-level learner persistence patterns identified in Cycle 1 into actionable strategic insights for course providers. Specifically, this cycle evaluates whether countries exhibiting low learner persistence may benefit from offline or blended learning interventions aimed at improving learner engagement and potential revenue.

Cycle 2: Data Understanding

Cycle 2 reuses the country-level persistence metrics derived in Cycle 1. No additional datasets are introduced. The analysis is based on aggregated learner persistence indicators, including average maximum step reached and early dropout rates by country.

Cycle 2: Modelling / Analysis

Segmentation Approach- Rather than predictive modelling, this cycle applies a rule-based segmentation approach to classify countries into persistence categories. Countries are categorised as low, medium, or high persistence based on interpretable thresholds applied to learner engagement and early dropout rates, supporting clear stakeholder decision-making.

Country Classification Countries were classified using quantile-based thresholds applied to average learner progression and early dropout rates.

```
country_cycle2 <- read_csv("../cache/country_cycle2_classification.csv")
```

```
## Rows: 190 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): country, persistence_category
## dbl (4): learners, avg_max_step, median_max_step, early_dropout_rate
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Cycle 2: Results and Visualisation

Learner Engagement by Country Figure X shows average learner progression by country, highlighting countries with low persistence that may benefit from targeted offline or blended learning interventions.

```
ggplot(
  country_cycle2,
  aes(x = reorder(country, avg_max_step), y = avg_max_step, fill = persistence_category)
) +
  geom_col() +
  coord_flip() +
  labs(
    title = "Average Learner Progression by Country",
    x = "Country",
    y = "Average Maximum Step Reached"
  )
```

Average Learner Progression by Country
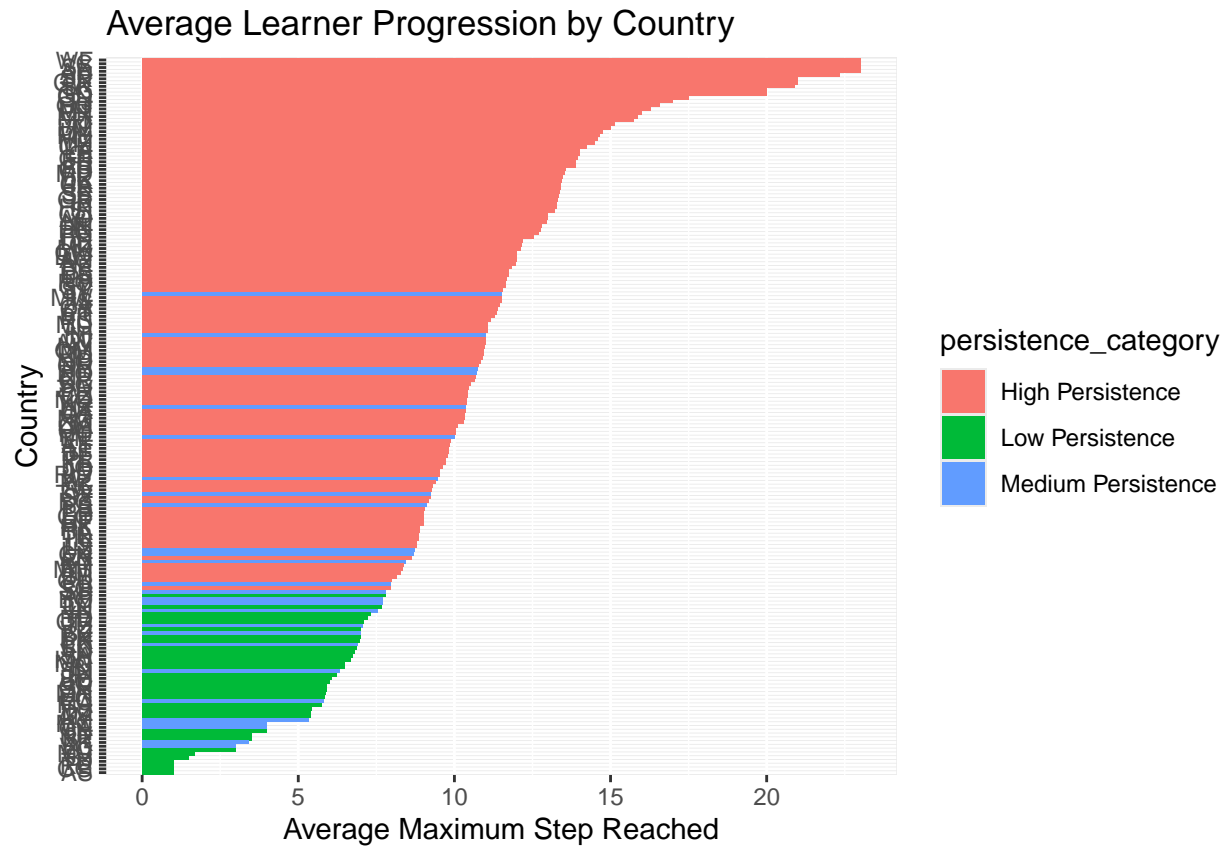
Early Dropout and Engagement Relationship Figure Y examines the relationship between learner engagement and early dropout rates across countries.

```
ggplot(
  country_cycle2,
  aes(x = avg_max_step, y = early_dropout_rate, color = persistence_category)
) +
  geom_point(size = 3) +
  labs(
    title = "Early Dropout Rate vs Learner Engagement",
    x = "Average Maximum Step Reached",
    y = "Early Dropout Rate"
  )
```

Early Dropout Rate vs Learner Engagement
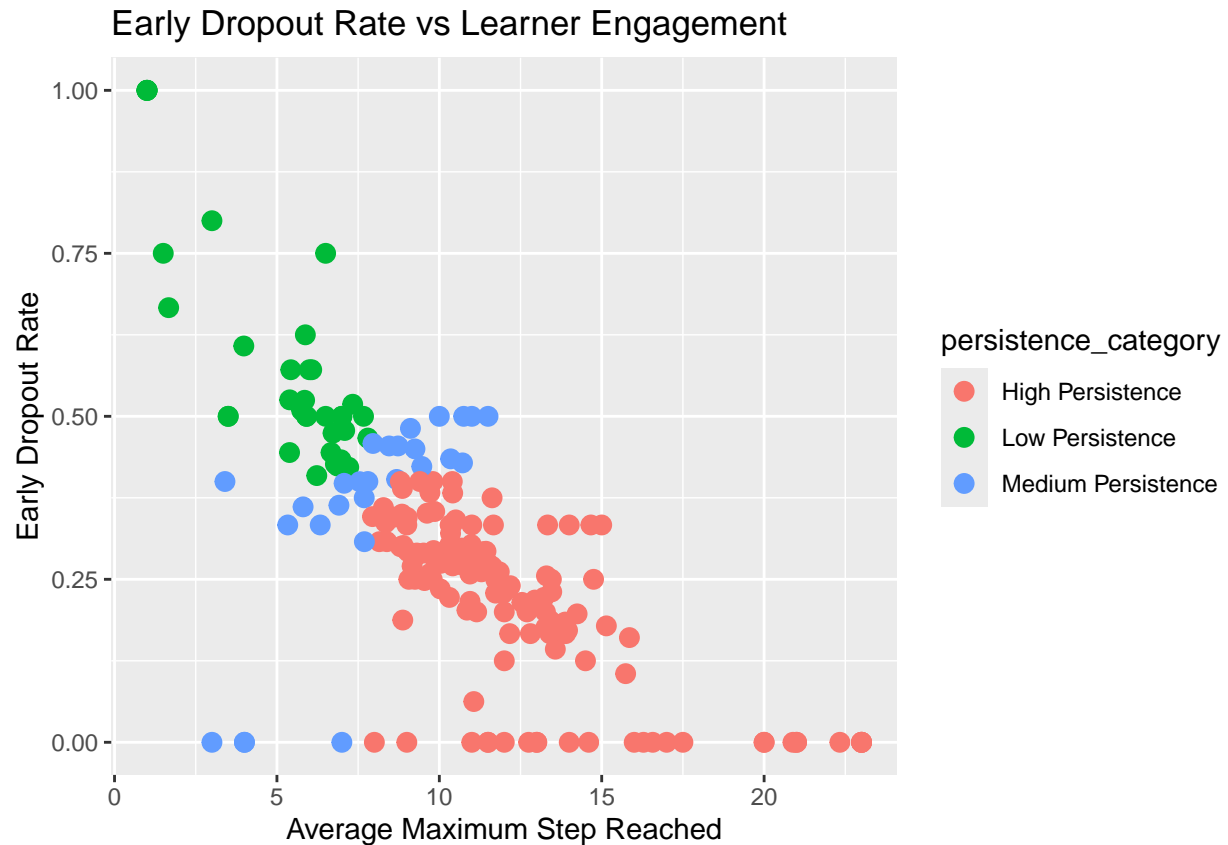
Low Persistence Countries Table Z lists countries classified as low persistence, representing potential candidates for targeted offline or blended learning interventions.

```
low_persistence_countries <- read_csv("../cache/low_persistence_countries.csv")
```

```
## Rows: 35 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): country, persistence_category
## dbl (4): learners, avg_max_step, median_max_step, early_dropout_rate
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
low_persistence_countries
```

```
## # A tibble: 35 x 6
##    country learners avg_max_step median_max_step early_dropout_rate
##    <chr>      <dbl>        <dbl>           <dbl>              <dbl>
## 1 AS             1            1               1                  1
## 2 CG             1            1               1                  1
## 3 RE             1            1               1                  1
## 4 TD             1            1               1                  1
## 5 LC             5            3               2                0.8
## 6 BI             4          6.5               1               0.75
```

```
##  7 BJ                4        1.5            1                0.75
##  8 MV                3        1.67           1                0.667
##  9 MN                8        5.88           2                0.625
## 10 TN               51        3.98           1                0.608
## # i 25 more rows
## # i 1 more variable: persistence_category <chr>
```

Cycle 2: Evaluation

Countries classified as low persistence exhibit both low average learner progression and high early dropout rates. These patterns suggest the presence of structural or contextual barriers rather than a lack of learner interest. Offline or blended learning models may reduce these barriers by improving accessibility, engagement, and course completion, thereby increasing the likelihood of certificate purchases and associated revenue.

Cycle 2: Deployment and Recommendations

It is recommended that course providers pilot offline or blended learning initiatives in selected low-persistence countries identified in this analysis. Cycle 1 persistence metrics should be retained as a baseline, and the same analytical framework reapplied post-intervention to evaluate changes in learner engagement, dropout rates, and potential revenue impact.