Cyber Security MOOC Data Analytics

Rushi Girdharbhai Vasoya

2026-01-15

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