Data Management and Exploratory Data Analysis - CSC8631| Analysing Engagement & Dropout in Newcastle University's Cybersecurity MOOC: Insights to Boost Learner Retention

Monisha Dabbara (Student ID: 240503817)

2024-11-10

Introduction

Futurelearn is a British open online course provider, founded in 2012 and partnering with the leading universities and organisations in the UK to provide massive open online courses, micro-credentials and full degrees. FutureLearn, originally created by The Open University, offers more than 2000 courses from around the world in business, technology, health, arts and more. Its user-friendly interface and focus on social learning through peer engagement and discussion boards to facilitate student engagement and learning with other users are the strengths of FutureLearn.

The provided data spans 7 years of a massive open online course (MOOC) developed by Newcastle University and run by the online skills provider FutureLearn. The course was titled "Cyber Security: Safety At Home, Online, and in Life" which ran for three weeks, and provided an option for learners to pay for a certificate of completion if desired.

Cycle 1 of CRISP-DM

1. Business Understanding

Business Understanding phase addresses both objectives and goals for analysing learner engagement and retention in the Newcastle University Cybersecurity MOOC on FutureLearn, we structure the project using the CRISP-DM framework. This approach will help clarify the steps needed to extract insights and provide actionable recommendations to enhance course retention.

1.1 What are the desired outputs of the project?

Demographic and Behavioural Factors Influencing Dropout

- Analyse demographic factors (age, gender, country) to find correlations with higher dropout rates and identify learner groups more likely to disengage to understand potential barriers to course completion.
- Examine behavioural patterns that may lead to dropout and stages associated with attrition, providing insights for improving course design and learner support.

Learner Engagement and Retention Patterns

- Explore the impact of engagement (content interaction, quiz completion) on course retention and completion rates and assess engagement metrics (time on modules, quiz scores, forum activity) to understand their influence on retention.

- Identify specific modules or course weeks with high dropout rates to detect content-related disengagement.

1.2 What Questions Are We Trying To Answer?

- To Determine demographics (age, gender, country) and behaviours correlated with high dropout rates.
- Identify key engagement metrics and content areas associated with retention, including drop-off points.
- Provide actionable insights and recommendations to reduce dropout and improve course completion rates.

2. Data Understanding

In this phase, I will evaluate the data requirements aligned with our objectives, assess the availability of pertinent data, and analyse the reliability of our sources.

2.1 Initial Data Report

As per the project data requirements, the raw data consists of seven years of learner interaction and progress data from a massive open online course (MOOC) developed by Newcastle University and collected by FutureLearn. This data is stored in the project's data folder in the form of raw .CSV files.

For this analysis, three key datasets are considered: enrolments, leaving-survey-responses, and step activity. However, leaving survey responses are only available for batches 4-7, so only 4-7 batches will be considered. Then, to integrate these datasets comprehensively, they are merged using a left join on the 'learner_id' field, creating a unified dataset for further analysis. The mentioned Preprocessing steps are done in '01-A.R' file in munge folder of this project.

2.2 Initial Exploration

To begin with, I will load the datasets and conduct an initial exploratory data analysis (EDA) to familiarise with their structure and content. This involves checking for missing values, confirming data types, and calculating basic statistics to understand the distributions of essential demographics (such as age, gender, and country), engagement metrics (like activity_step), and indicators of dropout behavior (indicated by is_dropout).

This process not only ensures data completeness and accuracy but also highlights key attributes that will guide the analysis. Alongside data type issues, missing values are a frequent challenge in real-world data. These may arise for various reasons and must be addressed—either by filling in or removing them—before training any machine learning models. This thorough examination sets the foundation for a cleaner dataset and more reliable insights in the next stages of analysis.

3. Data Preparation

In the **Data Cleansing** and **Data Wrangling** stages of the CRISP-DM framework, under the **Data Preparation** phase, the focus is on handling missing values and transforming raw data to ensure accuracy, reliability, and structure for analysis. Here's how each step fits into this process:

3.1 Data Cleansing

Objective: The goal is to address missing or inconsistent data to maintain both accuracy and completeness in the dataset.

Handling Missing Values: A critical step in data preparation involves addressing missing values, particularly in essential columns such as age range, gender, and country. Since these attributes are fundamental to the demographic analysis, records with missing values in these fields are removed. Excluding these incomplete records helps preserve data integrity, ensuring that subsequent analyses reflect an accurate and comprehensive view of the learner population.

By handling missing values early on, I can prevent potential biases and inaccuracies that might otherwise skew insights, particularly in demographics-based analyses. This proactive approach guarantees that key findings on age, gender, and geographical representation are both reliable and representative, laying a strong foundation for the analysis.

Defining Dropout Status:

A new variable, “dropout status” is introduced to mark whether a learner dropped out of the course or continued. This variable is binary, with a value of 1 indicating a dropout and 0 indicating continuation. The dropout status is determined based on the “left_at” field: learners with a value in this field are considered to have dropped out.

Defining dropout status allows for precise measurement of retention rates and a clearer identification of patterns linked to dropout behavior. By establishing this variable, I can better understand which factors may influence learner retention, providing actionable insights to improve course engagement and reduce dropout rates.

3.2 Data Wrangling

Objective: Transform and organise the data for analysis.

Combining Data Sources:

- Data from multiple datasets—including Enrolments, Leaving Survey, and Step Activity—are merged. Merging these sources ensures a comprehensive view of each learner, including demographic details, engagement levels, and dropout indicators.
- Combining these datasets allows for a holistic perspective on each learner, which is necessary for identifying relationships between engagement, demographics, and retention.

Aggregating by Demographics and Engagement:

- Once the data is cleaned, it is aggregated by age group, gender, and country to calculate average dropout rates across these demographics.
- Aggregation enables targeted insights into how specific demographic groups or engagement patterns impact dropout rates, helping to identify where retention efforts may need to be focused.

Outcome of Data Cleansing and Wrangling

These data preparation steps ensure the dataset is complete, consistent, and ready for analysis. This approach maximises the accuracy of insights into dropout patterns, making it possible to explore how demographic factors influence course retention and identify targeted strategies for improvement.

To recap, handled missing values in the combined dataset and introduced new variables (dropout_status) for further analysis of the data.

4. Modelling

In the **Data Modelling** phase of CRISP-DM, the focus is on creating statistical summaries and models to identify patterns and relationships within the data. This phase uses calculations and aggregations to provide insights that help address business objectives—in this case, understanding dropout rates and engagement patterns across different demographics.

4.1 Understanding Dropout Rates for Demographic Groups

To better understand dropout trends, data is grouped by demographic categories—such as age group, gender, and country. For each demographic group, the dropout rate is calculated by finding the average value of the dropout indicator (where dropout is represented by 1 and continuation by 0). This calculation reveals the proportion of learners who dropped out within each demographic segment.

By examining dropout rates across different groups, it is possible to identify specific demographics where dropout is more prevalent, enabling targeted interventions. For example, if younger age groups show higher dropout rates, adjustments in content or support for that age range might improve retention.

4.2 Descriptive Analysis

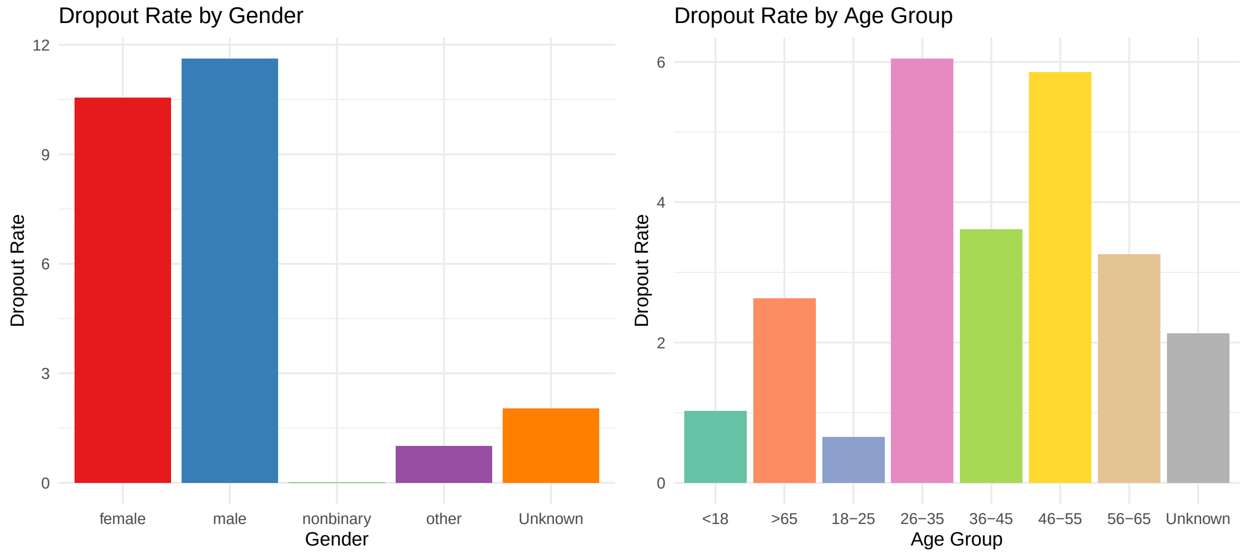
Completion Rates by Demographic Group: In addition to dropout rates, completion rates across demographics (age and gender) are calculated to get a balanced view of course engagement. By calculating completion rates for each demographic, this analysis reveals which groups are most likely to complete the course. These insights can guide decisions on how to retain and support learners from groups with lower completion rates.

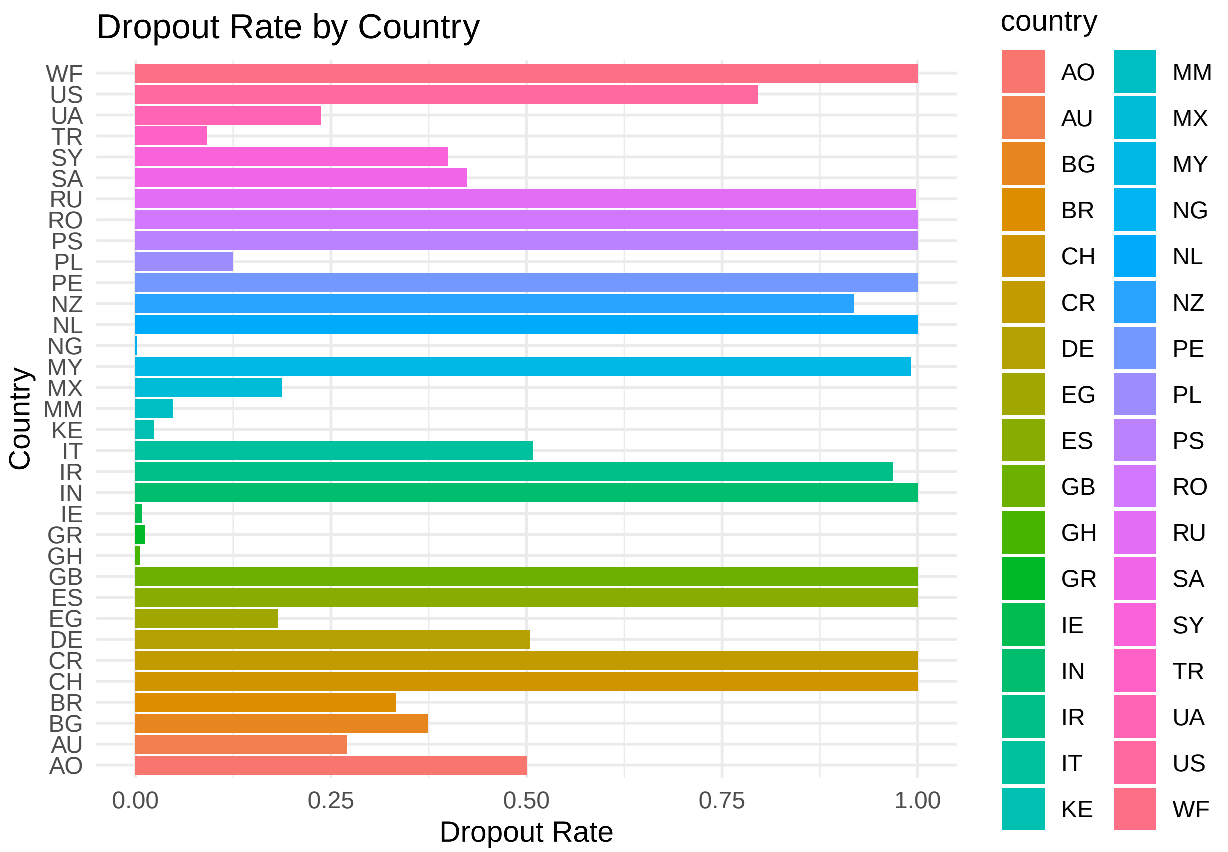
Engagement Metrics: Analysing additional engagement metrics—such as the average time spent on each step and the number of steps completed—offers a more nuanced understanding of learner behaviours. This involves calculating:

- **Average Time Spent**: Determining the typical amount of time learners invest per step or module can indicate engagement levels. Low time investment may suggest content is too challenging, too easy, or not engaging enough.
- **Steps Completed**: The average number of steps completed across demographics shows the extent to which different groups engage with the course material, helping to identify at which stages or content types learners tend to disengage.

Together, dropout rates, completion rates, and engagement metrics provide a comprehensive view of the course's impact on various demographics. This modelling phase transforms raw data into actionable insights, highlighting patterns and trends that inform strategies for improving learner retention and engagement.

The Visualisation below displays dropout rates by demographic groups to spot patterns.





Note: Excluded the countries with '0' dropouts.

5. Evaluation

In the **Evaluation** phase, I evaluated the quality of the findings against the initial business objectives. Key insights reveal trends in dropout rates by demographics, highlighting areas for improvement.

Countries with the highest dropout rates include **PS, RO, PE, CR, ES, CH, NL, WF, GB, and IN**. This suggests that factors such as language barriers, internet access, or cultural relevance may significantly impact learner engagement. Additionally, the analysis indicates that **males** exhibit the highest dropout rates, implying that the course content or structure may not engage male learners effectively.

Further analysis shows that the **26–35** age group has the highest dropout rate, followed by the **46–55** age group. This finding suggests that younger learners may face time constraints due to personal and professional commitments, while older learners might struggle with the accessibility of course content. Identifying specific weeks or modules with high dropout rates can help pinpoint areas with challenging content or low engagement activities.

For the **26-35** age group, providing flexible pacing and reminders can help accommodate their schedules. Lastly, leveraging dropout data to refine content difficulty and adding supportive materials for specific modules will further enhance the learning experience.

In conclusion, the insights gained from this analysis align with the objective of understanding dropout patterns and demographic influences on retention. They provide a strong foundation for actionable recommendations to improve course design. Future analyses could delve into more granular engagement metrics to deepen the understanding of learner behavior.

Cycle 2 of CRISP-DM

Using insights from the first cycle, let’s deepen the analysis to address the steps where dropouts occur.

1. Business Understanding (Refinement)

In the first cycle of business understanding, I analysed dropout rates based on demographic factors such as age, gender, and country. However, to effectively enhance the course, it is essential to identify the specific stages or steps within the course where significant numbers of learners are dropping off.

This focused investigation aims to obtain the exact activity steps that exhibit high dropout rates, allowing for a more granular understanding of where learners are facing challenges. By segmenting this analysis by age group, gender, and country, one can uncover patterns that may indicate why certain demographics struggle at specific points in the course.

Objective Refinement: Investigate specific activity steps within the course where dropout rates peak, by age group, gender, and country

2. Data Understanding

In the first cycle of CRISP-DM, the focus was on understanding the enrolment data and leaving survey responses to gather initial insights into learner demographics and reasons for dropping out. In the second cycle, the emphasis shifts to the Step Activity dataset to investigate at which specific stages of the course learners are most likely to drop out.

To achieve this, I began by summarising dropout occurrences by activity_step along with demographic variables such as age range, gender, and country. This summary provides a comprehensive overview of dropout rates across different stages of the course, allowing for targeted analysis of where learners are disengaging.

The analysis involves grouping the data by the step within the course, as well as by demographic factors. By calculating the dropout rate for each combination of step, age range, gender, and country, patterns can be identified to reveal which specific activities may be challenging for learners.

3. Data Preparation

In preparing the data for this analysis, I filtered the dataset to focus solely on entries where learners have unenrolled from the course. This ensures that the analysis is centred on actual dropout events. Additionally, I cleaned and transformed the data by removing any entries with “Unknown” values for age or gender, as these would not provide meaningful insights into demographic trends.

By concentrating on the Step Activity dataset, I can gain deeper insights into learner behaviour and identify critical points in the course where interventions could be implemented to improve retention. This focused analysis will facilitate a better understanding of the specific challenges faced by learners at different stages, ultimately guiding the development of strategies to enhance course engagement and completion rates.

4. Modelling

In the **Modelling** phase of the second cycle of CRISP-DM, I focused on recalculating dropout rates by specific activity steps within the course. This analysis is crucial for understanding at which stages learners are most likely to disengage and ultimately drop out. By examining these dropout rates, I can identify patterns that may indicate issues with course content or engagement strategies.

To begin, I grouped the dataset by **activity step** and calculated the overall dropout rate for each step. This summary allows for a straightforward identification of steps where dropout rates are particularly high.

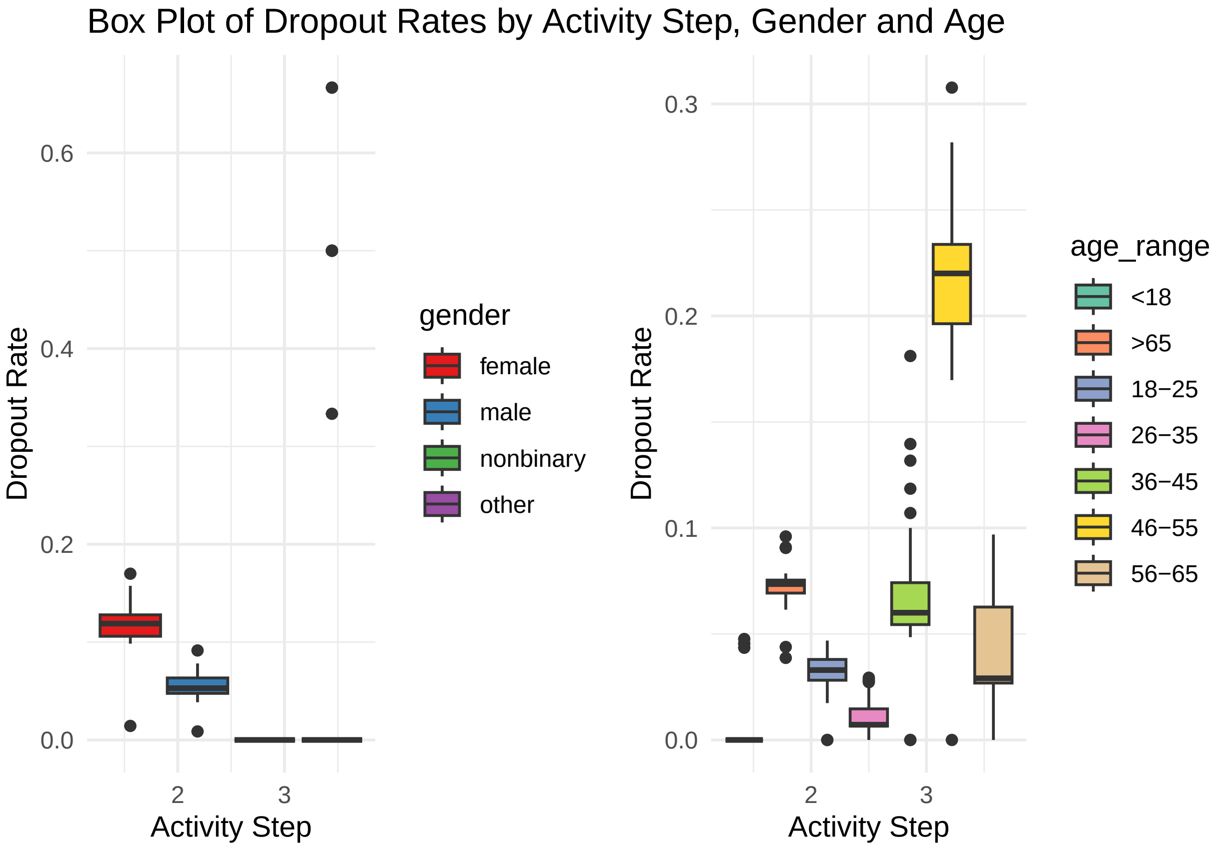
Summary of ‘activity step’:

step	dropout_rate
Min. :1.100	Min. :0.007173
1st Qu.:1.625	1st Qu.:0.039676
Median :2.205	Median :0.040949
Mean :2.335	Mean :0.040344
3rd Qu.:3.138	3rd Qu.:0.042032
Max. :3.900	Max. :0.053004
NA’s :1	NA

I expanded the analysis by incorporating demographic factors. I filtered the dataset to exclude any entries with “Unknown” values for age range and gender. I then grouped the data by **step** and **gender** to calculate dropout rates specifically for each gender at each activity step. Similarly, I performed an analysis by **step** and **age range** to observe how dropout rates varied across different age groups. This dual approach enables a nuanced understanding of how different demographics interact with course content.

The comparative analysis also includes plans to examine engagement metrics between learners who completed the course and those who left at different stages.

Additionally, I summarised dropout counts by age and gender, providing further context to the dropout rates. This summary reveals the absolute number of dropouts within each demographic group, highlighting which groups are more affected by course attrition.



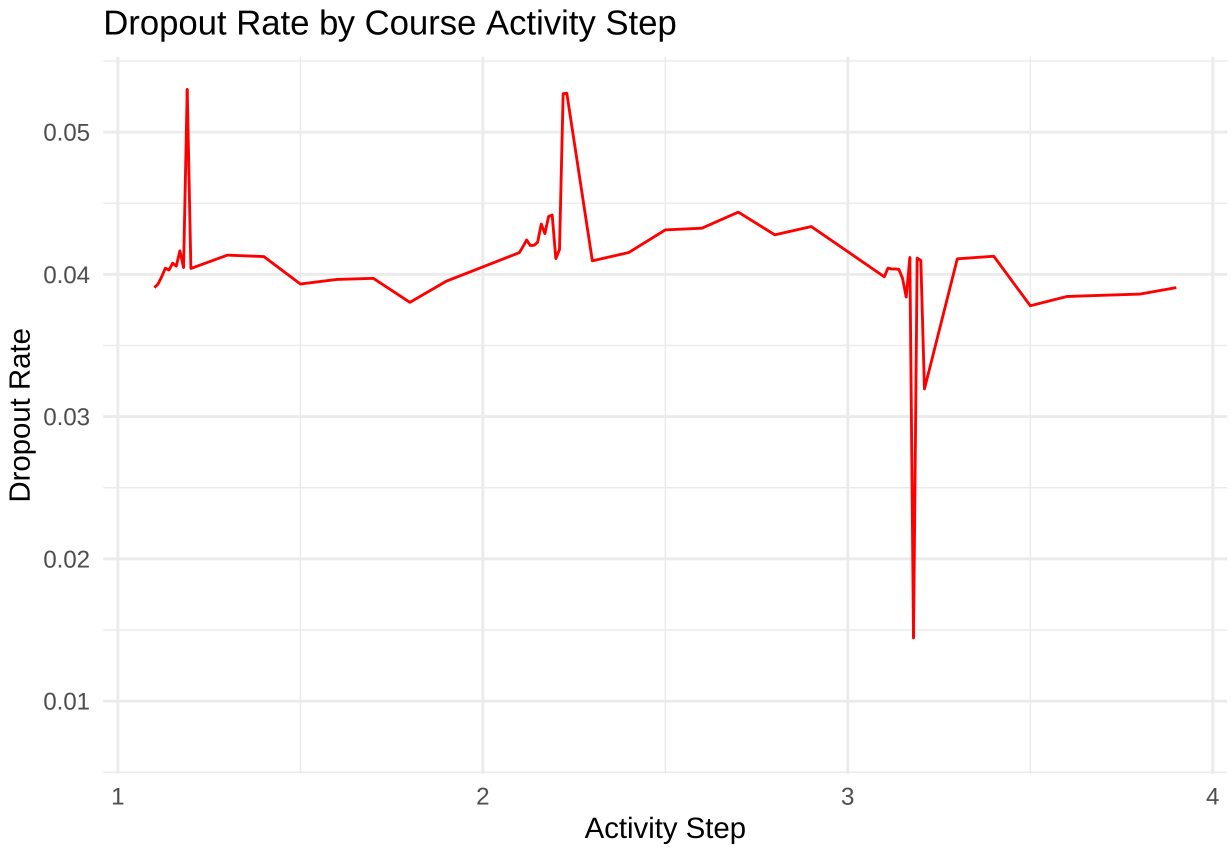
For visualisation, I created box plots to represent dropout rates by activity step as shown above, differentiating the data by gender in one plot and by age range in another. The box plots provide a clear visual representation of dropout rate distributions across different activity steps, allowing for easy comparison of engagement metrics between genders and age groups.

These advanced visualisations and statistical comparisons will enhance the modelling phase, allowing for deeper insights into learner behaviour and the factors influencing dropout rates. This comprehensive approach will ultimately inform targeted recommendations for improving course design and engagement strategies.

5. Evaluation

In the second cycle of the CRISP-DM framework, the evaluation phase focuses on analysing the dropout patterns identified in the data and assessing how these insights can inform course improvements. The visualisations produced during this cycle highlight critical trends regarding learner engagement and retention at different activity steps.

Plot dropout rate by activity step

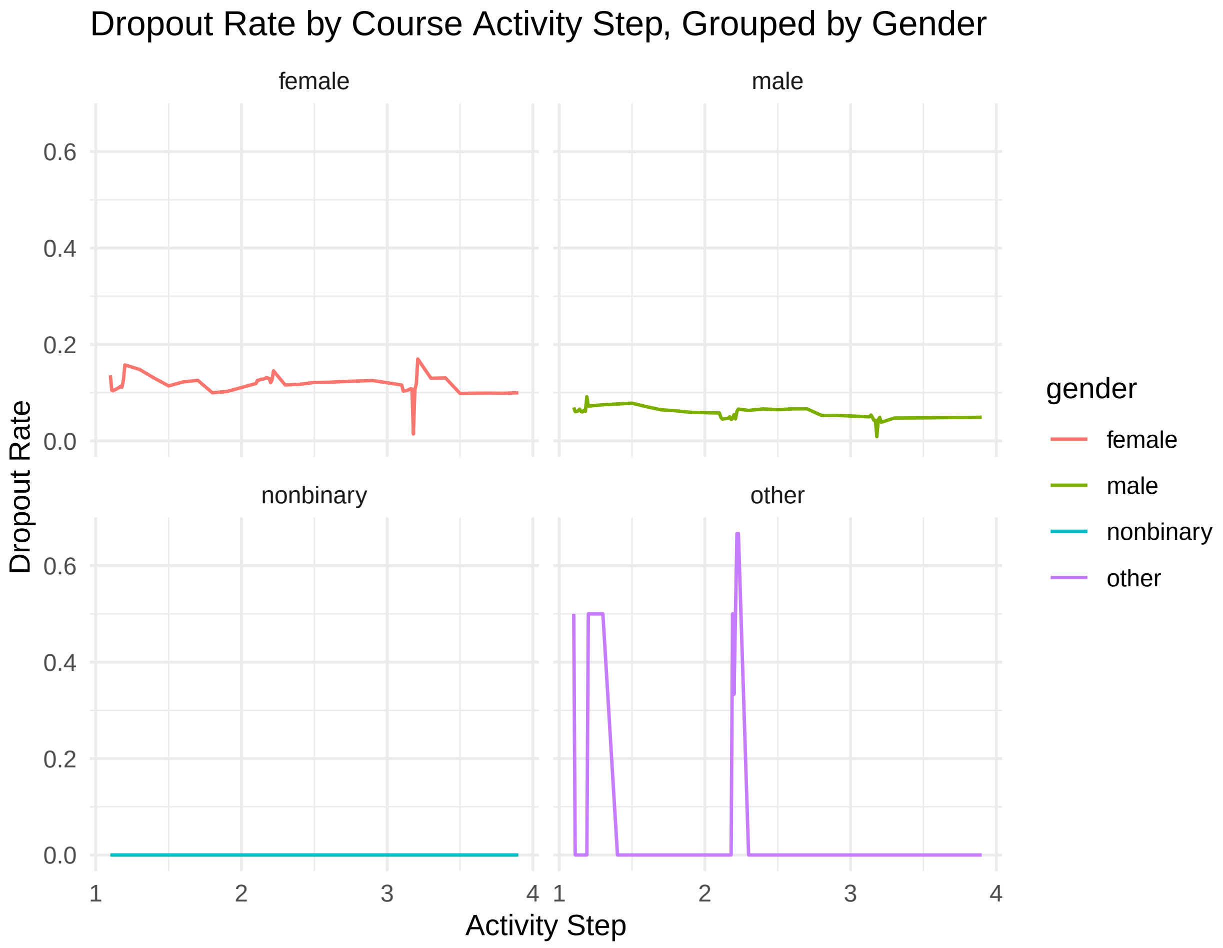


Visualisation of Dropout Rates Across Activity Steps

The line plot of dropout rates by course activity step shown above reveals crucial insights for optimising learner retention at different stages. Notably, dropout rates peak around the 1.25 and 2-2.5 steps, suggesting learners encounter challenges during these early stages, which might correlate with common issues in online learning like perceived isolation, lack of engagement, or difficult course material. Research highlights that disengagement in online courses often results from a combination of factors: lack of community, time conflicts, and ineffective course design, which could contribute to dropouts at these early steps.

On the other hand, the dropout rate is lowest at the 3-3.5 activity step, suggesting that learners who reach this phase might have adjusted to the course structure and content, leading to better engagement. This pattern aligns with findings that once students navigate early challenges, they are more likely to persist. Therefore, targeted curriculum adjustments or additional support at the initial stages could be beneficial in reducing dropout rates.

To plot the dropout rate by activity step by gender



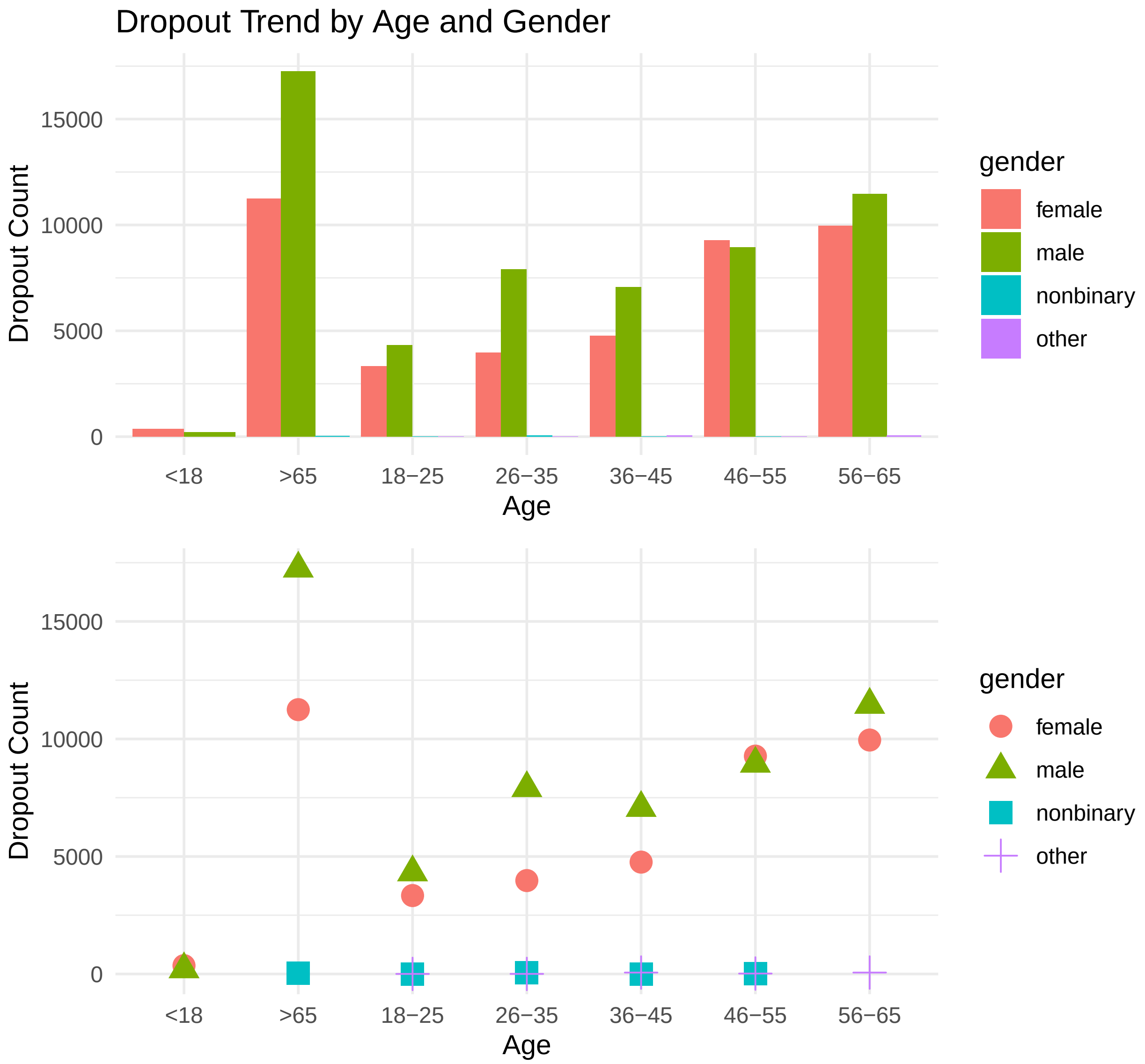
Gender and Age Trends in Dropout Rates

In the gender-based analysis, males show higher dropout rates across most activity steps, with further details indicating that learners over 65 experience the highest dropout rates among age groups. This aligns with research that older learners may face unique obstacles, including less familiarity with technology or more complex external responsibilities, which can hinder engagement. Additionally, females in some studies are found to engage more consistently due to higher perceived community connections, suggesting that enhancing interactive elements could help retain male learners.

Gender-Based Dropout Trends:

The subsequent line plot as shown above, separates dropout rates by gender, provides deeper insights into how dropout patterns differ across male and female learners. This visualisation allows for a comparative analysis of dropout rates at various activity steps, identifying specific steps where male learners may require additional support. The faceting by gender further clarifies these trends, enabling targeted interventions.

To improve retention, consider implementing additional support mechanisms like onboarding resources or interactive modules to help learners establish a strong start. Given that curriculum redesign around steps 1.25 and 2-2.5 could address peak dropout phases, course content and interactivity might also benefit from refinement to maintain engagement across demographics. These adjustments aim to improve learners' journey from the start, potentially reducing dropouts and enhancing overall course satisfaction.



Age Group Analysis:
The scatter plot visualising dropout counts by age group and gender reveals that the highest dropout rates occur among males over the age of 65. This insight emphasises the importance of understanding demographic factors when analysing dropout patterns. By identifying vulnerable groups, course designers can develop tailored strategies to retain these learners.

Refined Insights

The second cycle has illuminated new trends and patterns in engagement and retention that were not evident in the initial analysis. The focus on specific activity steps and the breakdown by gender and age group provides actionable insights that can inform course design. For example, recognising that certain demographics are more prone to dropout at specific stages allows for a more nuanced approach to course improvements.

Updated Recommendations

Based on the refined findings, several recommendations emerge:

1. **Curriculum Adjustments**: Modify the content and structure at the 1.25 and 2-2.5 activity steps to address the identified peaks in dropout rates. This could involve enhancing instructional materials, introducing interactive elements, or providing additional resources to support learners.
2. **Targeted Support for Demographics**: Implement targeted interventions for male learners, particularly those over 65, who exhibit higher dropout rates. This may include personalised outreach, mentoring, or tailored content to better engage this group.
3. **Continuous Monitoring**: Establish ongoing evaluation mechanisms to track the effectiveness of implemented changes, allowing for iterative improvements based on learner feedback and engagement metrics.

In conclusion, the evaluation of dropout patterns during the second cycle of the CRISP-DM framework has provided a deeper understanding of learner engagement. The insights derived from the analysis will serve as a foundation for enhancing course design and fostering higher retention rates among diverse learner demographics.

6. Deployment

With insights gained from both cycles of analysis, the focus for deployment is now on implementing tailored recommendations to minimise dropout rates and enhance learner engagement, specifically addressing key points identified across demographic and activity stages.

1. Demographic-Specific Targeting:
Based on the analysis, learners over 65, male participants, and individuals from countries with high dropout rates were more likely to discontinue the course. Early-stage interventions for these groups can reduce dropout. Suggested strategies include:

- **Localised Support**: Offer region-specific resources, such as translated content, cultural examples, and targeted communication channels to meet localised needs.
- **Engagement Tools**: Since male participants tend to disengage at a higher rate, integrating more interactive elements (such as discussion forums, group activities, and challenges) could foster a sense of connection, shown to enhance male engagement in online learning environments.

2. Content Enhancement for High Dropout Steps:
Dropout rates spike around the 1.25 and 2-2.5 activity steps, indicating potential areas of complexity or disengagement. Based on feedback from the first cycle, refined recommendations include:

- **Content Simplification and Modular Support**: Introduce optional resources like videos or summaries at these steps to help learners grasp complex material. Content broken down into smaller, manageable modules can further reduce cognitive load and keep learners progressing.

- **Interactive Elements**: Including quizzes, live sessions, or real-world applications at these stages can make the content more engaging and clarify challenging concepts.

3. Targeted Engagement Strategies
To support learners across different demographics, targeted engagement strategies should be implemented:

- **Personalised Communication**: Automated email reminders or check-ins at critical points of the course can provide encouragement. Highlighting progress, achievements, or upcoming resources is known to enhance motivation.
- **Optional Cohort-Based Learning**: Facilitating group discussions for specific demographics (e.g., younger vs. older learners, industry newcomers) can encourage peer learning and accountability. In online courses, learners often benefit from cohort-based progression where they can discuss shared experiences and provide mutual support.

Refinement Based on First-Cycle Feedback
Feedback from the initial cycle of the CRISP-DM process emphasised the importance of adjusting content and support based on demographic trends. Following this, the refined recommendations in Cycle 2 now target both dropout-prone activity steps and specific demographic needs, focusing on personalised support and content accessibility enhancements for an optimised learning experience.

Final Recommendations Summary:

- **Demographic Targeting**: Identify and support at-risk demographics early in the course with localised and personalised support.
- **Content Enhancement**: Simplify or support content around high dropout steps and integrate engaging activities to maintain attention.
- **Engagement**: Personalised reminders and cohort-based learning options can improve retention by fostering motivation and community.

These deployment strategies, informed by patterns in learner data, set the groundwork for a structured approach to minimising dropout and improving course completion rates.