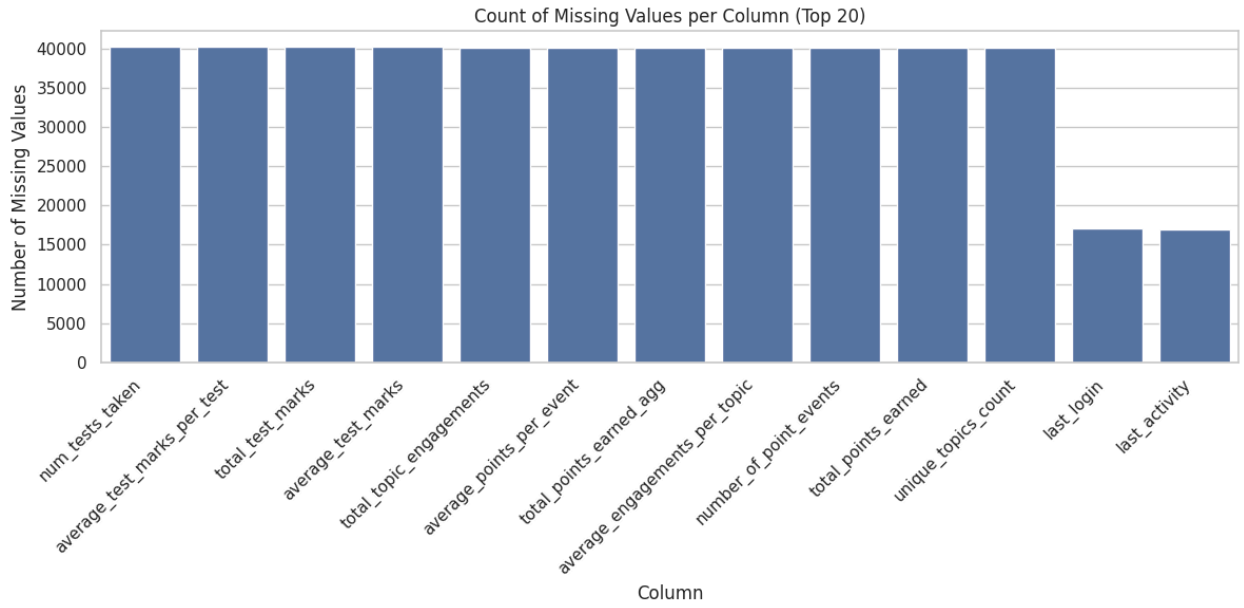
Summary of Key EDA Findings

Based on the exploratory data analysis conducted on eCampus test data:
This analysis was conducted on data provided from the test database with a secure connection. No data from the mix panel was provided. The data collected was very sparse and limited.

Data Availability and Engagement:

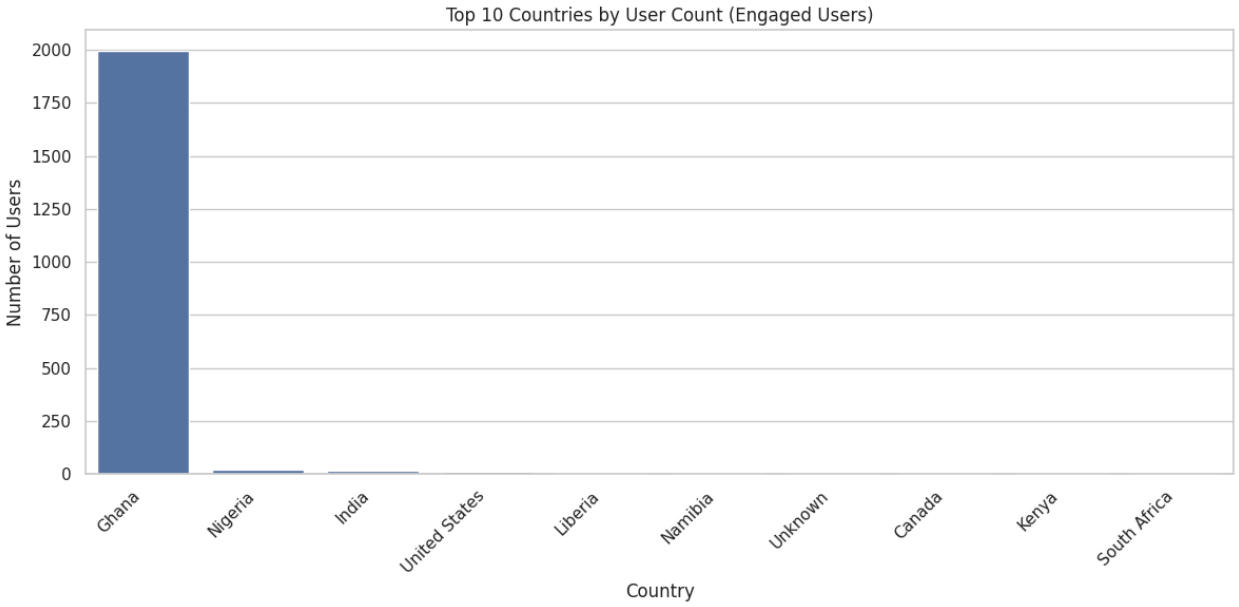
A significant portion of the over **42,000 users** in the test dataset do not have recorded activity, suggesting potentially low engagement in this test environment or data collection gaps.



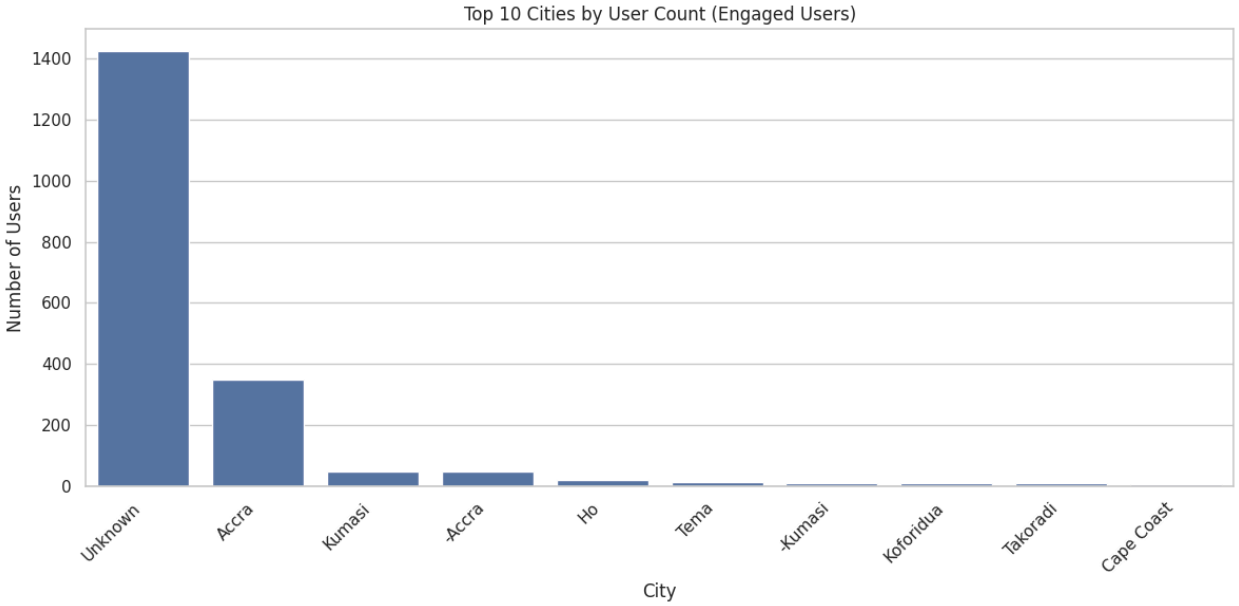
Focusing the analysis on 'engaged users' (those with at least 1 point earned) reduced the dataset significantly to around **2,100 users,** indicating the sparsity of engagement data in the current sample. ***The EDA was conducted on the engaged users only 5% of the entire data.***

User Demographics and Location:

The majority of engaged users are located in Ghana, followed by Nigeria and India.
A substantial number of users have 'Unknown' values for their city, highlighting a need for improved data collection in this area.

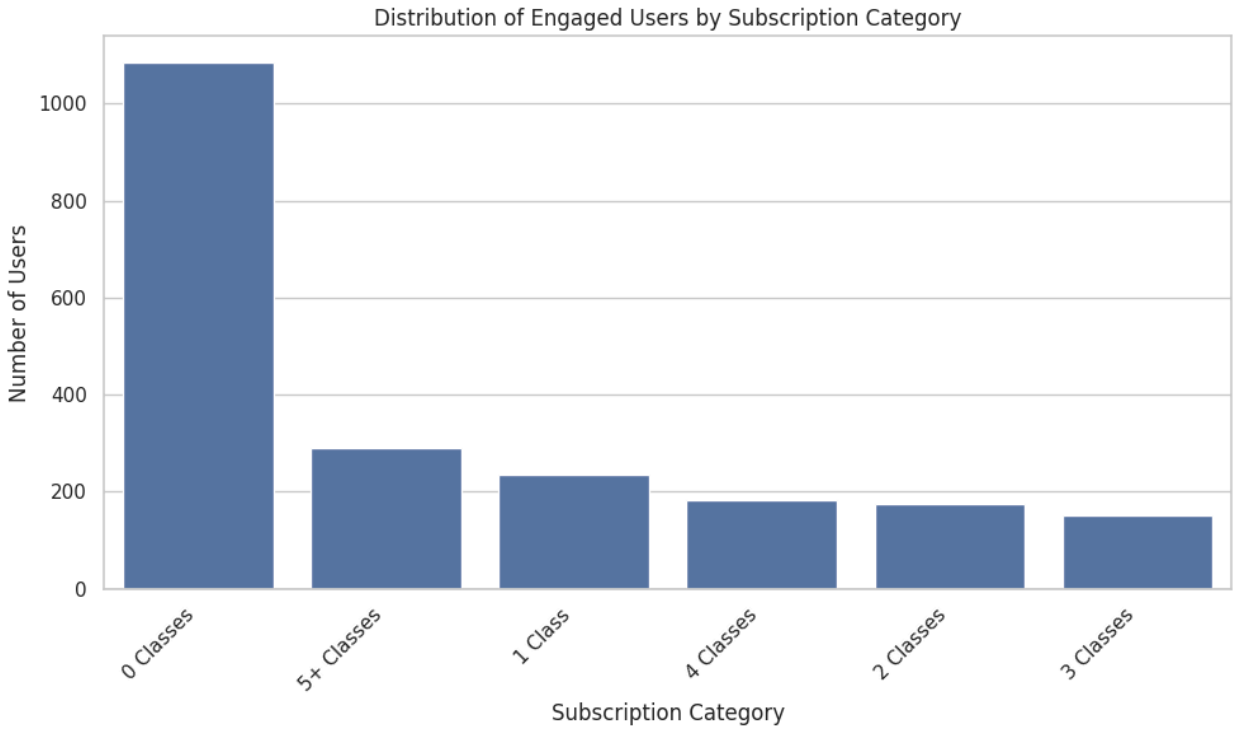


Case variations in city names were observed and standardized.

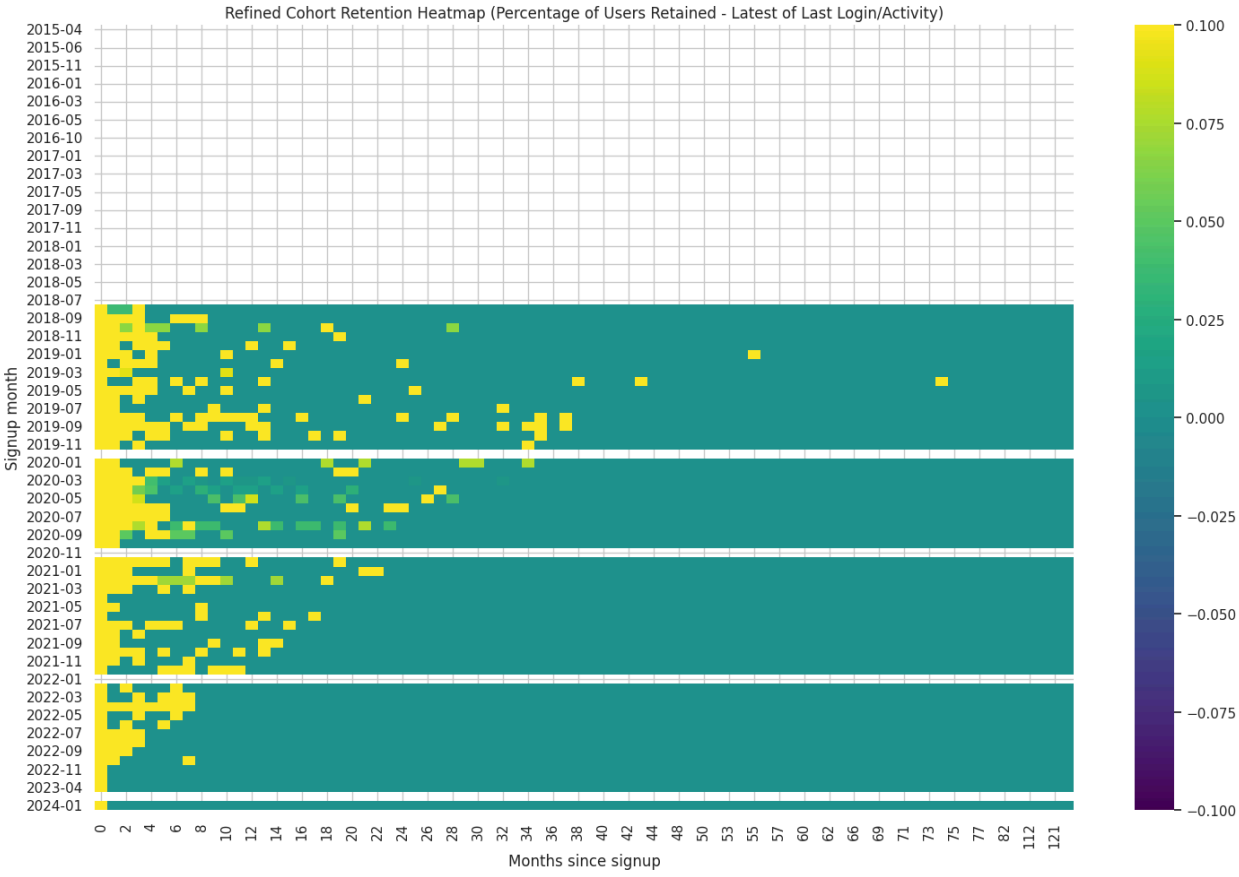


Users are not categorized by education level (shs/jhs etc) or based on gender or age. We could further get more insights with this data in future.

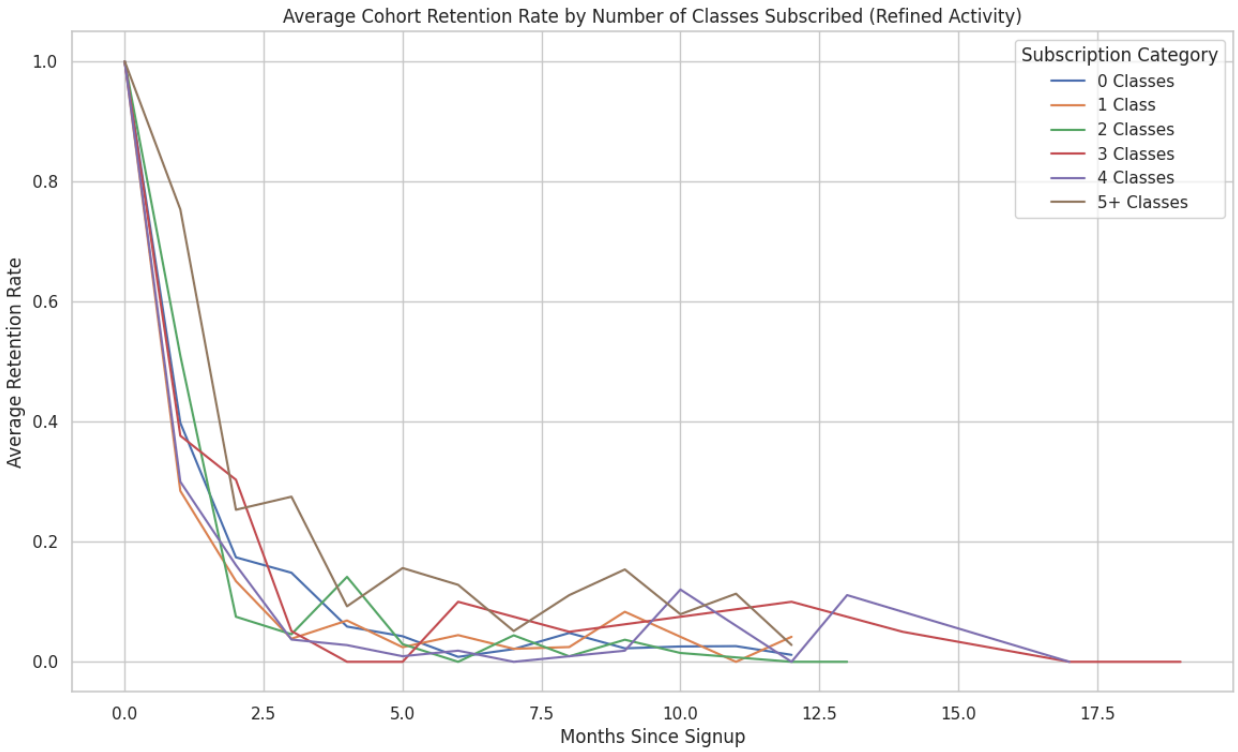
Retention Analysis:
Cohort analysis using 'last_login' showed rapid retention drop-off and significant missing data in this column, particularly for users with '0 Classes' subscribed.



A refined retention analysis using the latest of 'last_login' or 'last_activity' as the activity timestamp provided a broader view of user activity and showed retention curves for users with more subscribed classes.



Users who subscribed to more classes demonstrate significantly higher average retention rates compared to those with fewer or no subscribed classes. This is a critical business insight for encouraging subscriptions early in the user journey.



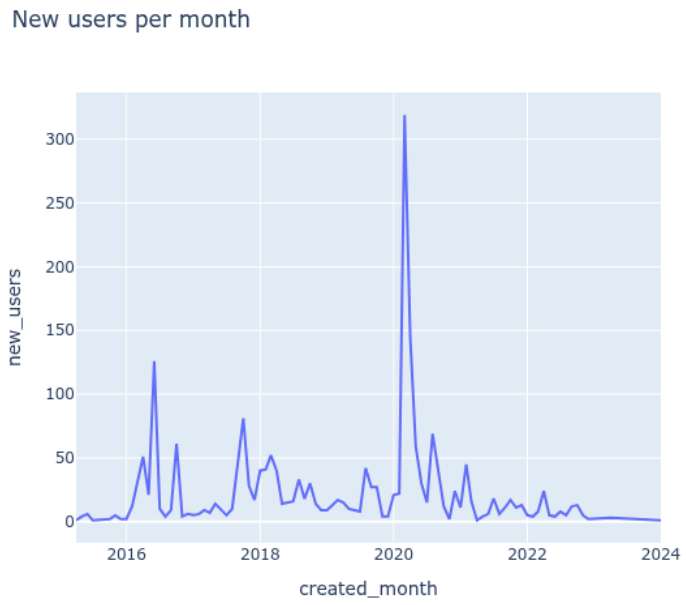
Analysis of retention based on user activity showed;

- ➔ 1 day Retention Rate: 13.83%
- ➔ 7 day Retention Rate: 30.83%
- ➔ 30 day Retention Rate: 41.97%
- ➔ 90 day Retention Rate: 46.13%
- ➔ 180 day Retention Rate: 50.99%
- ➔ 1 year Retention Rate: 54.67%

This means that with the engaged users, more than half of them were active throughout the year with a daily retention activity of ~14%.

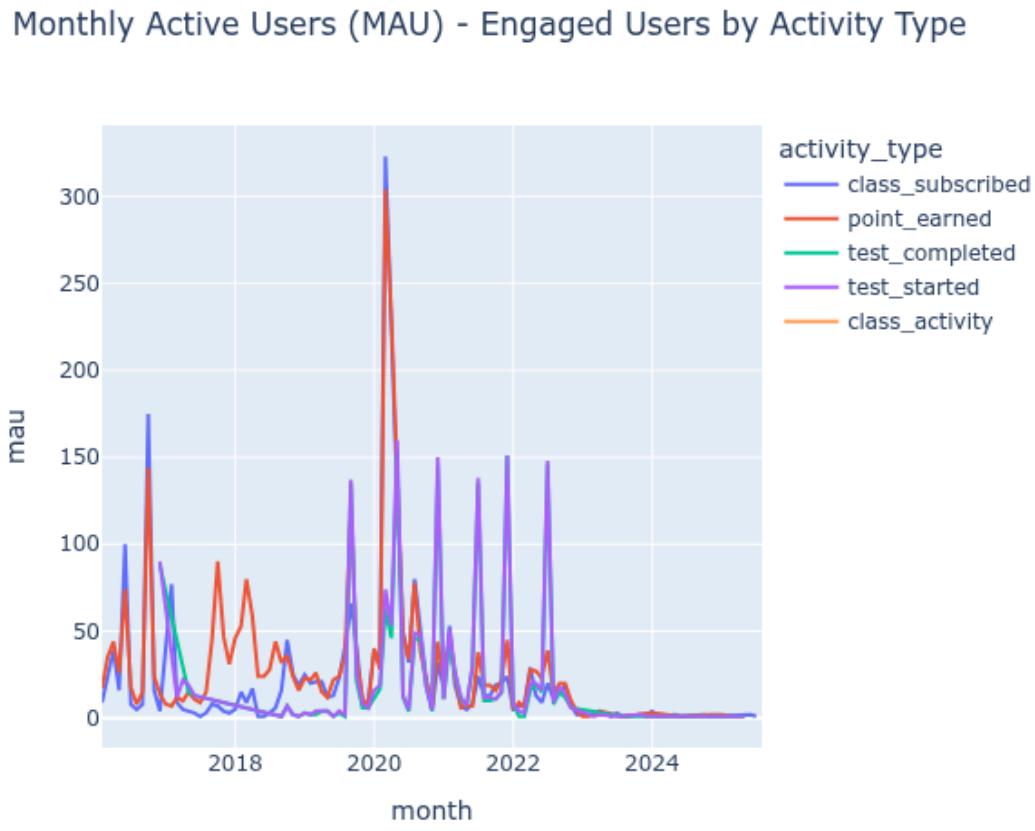
Time Series & Growth

Analysis of user account generation with engaged user dataset shows in as the peak of account creation. there anything specific that happened?



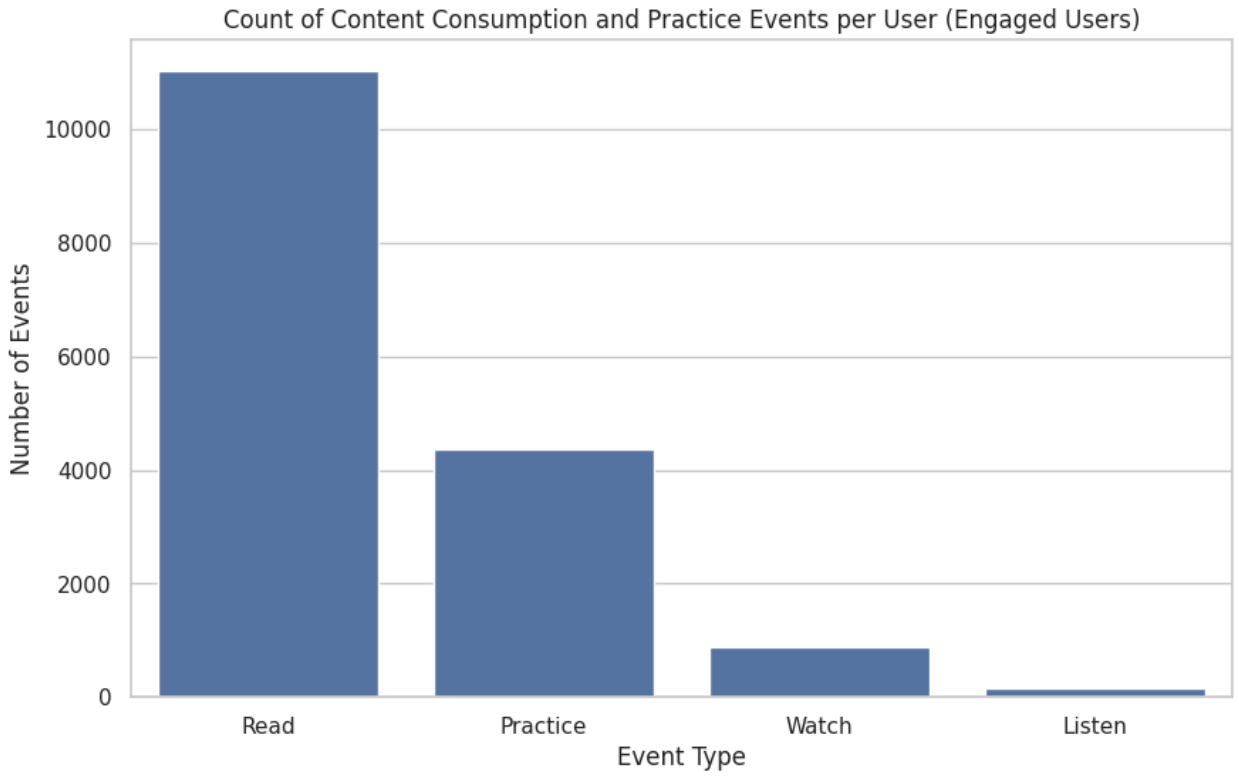
2020 Was

Further analysis on user activity (below) can show that lots of tests were started and not completed. A simple nudge to the user to encourage test completion will be handy here. We also see far less class activity like discuss, read, watch and listen.



Engagement Depth and Content Consumption:

Analysis of content consumption (Read, Watch, Listen) from points data showed distributions of event counts and marks earned, indicating varying levels and types of content interaction among engaged users.



`Here's a quick look at the content we have available:`

- Total content runtime: 27,705 minutes
- Practice content items: 38,331
- Test content questions: 29,604
- Total classes :135
- Total Content : 2632
- Class Categories : 8

A further analysis can be done with mixpanel to find out more on content consumption with minutes watched/listened to etc.

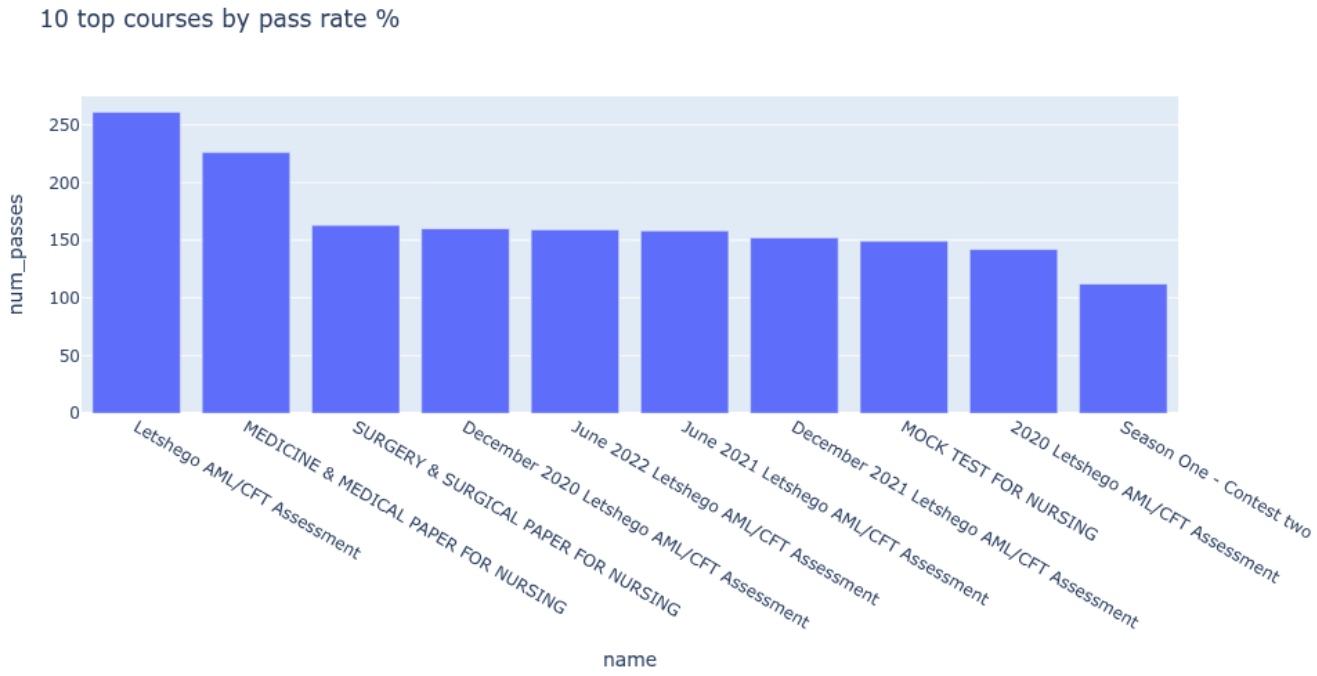
More research to be done into why watch and listen counts are really low, is it due to content availability?

Practice and Test Performance:

Practice points analysis revealed metrics like the number of practice events, average practice marks, and total practice marks.

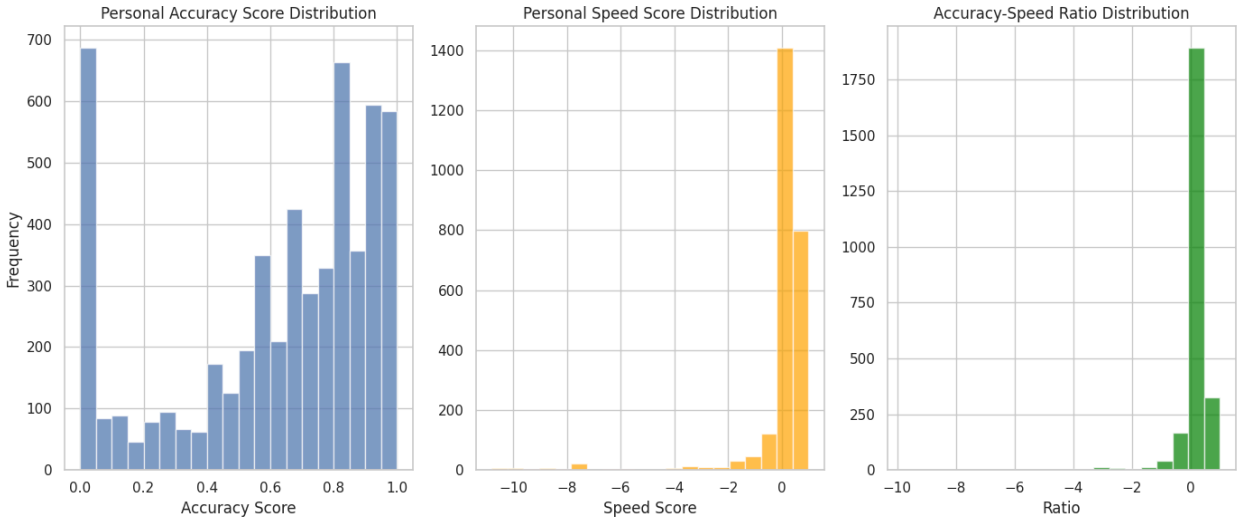
Analysis of test performance using test_takers and tests data provided insights into average test scores and pass rates for different tests among engaged users.

Below is a bar chart of top tests based on user pass rates, we can further compare these pass results to practice and see whether there is a correlation.

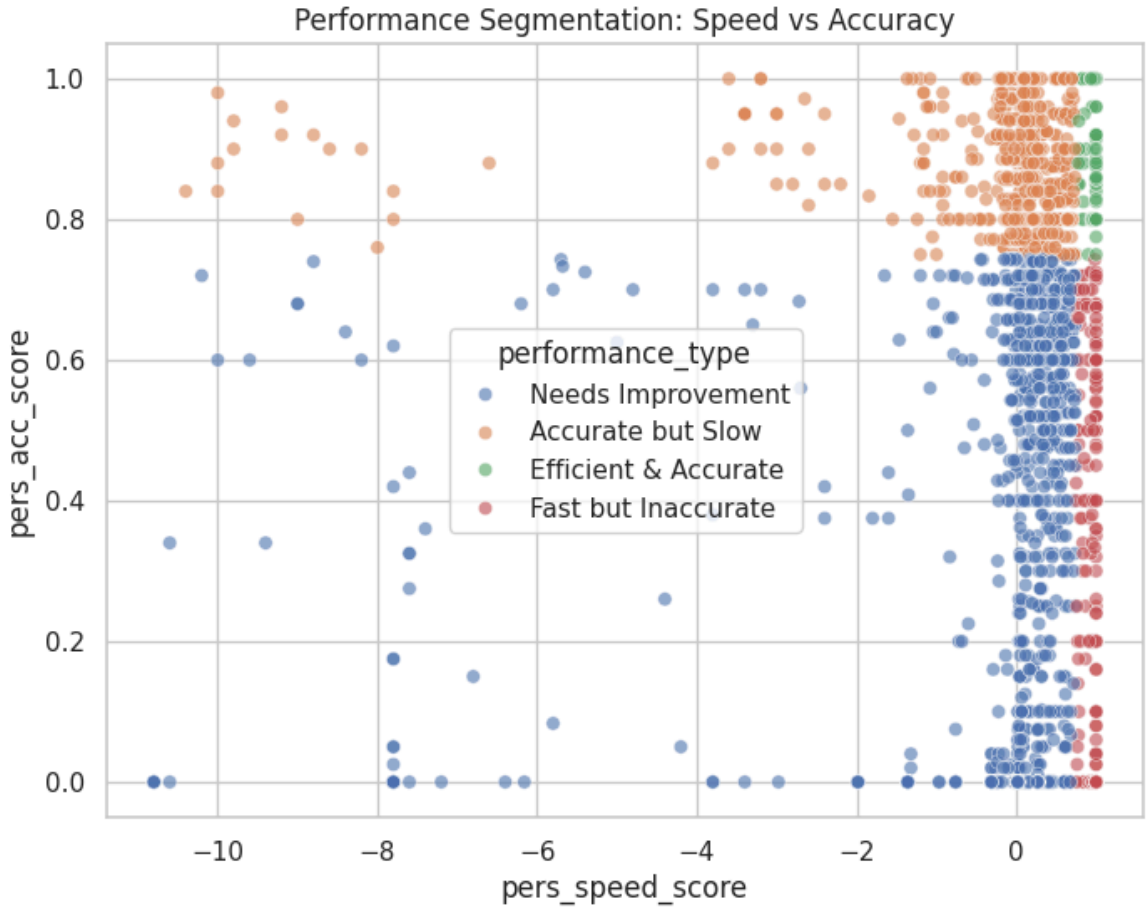


Accuracy to Speed Ratio (A:S) for tests.

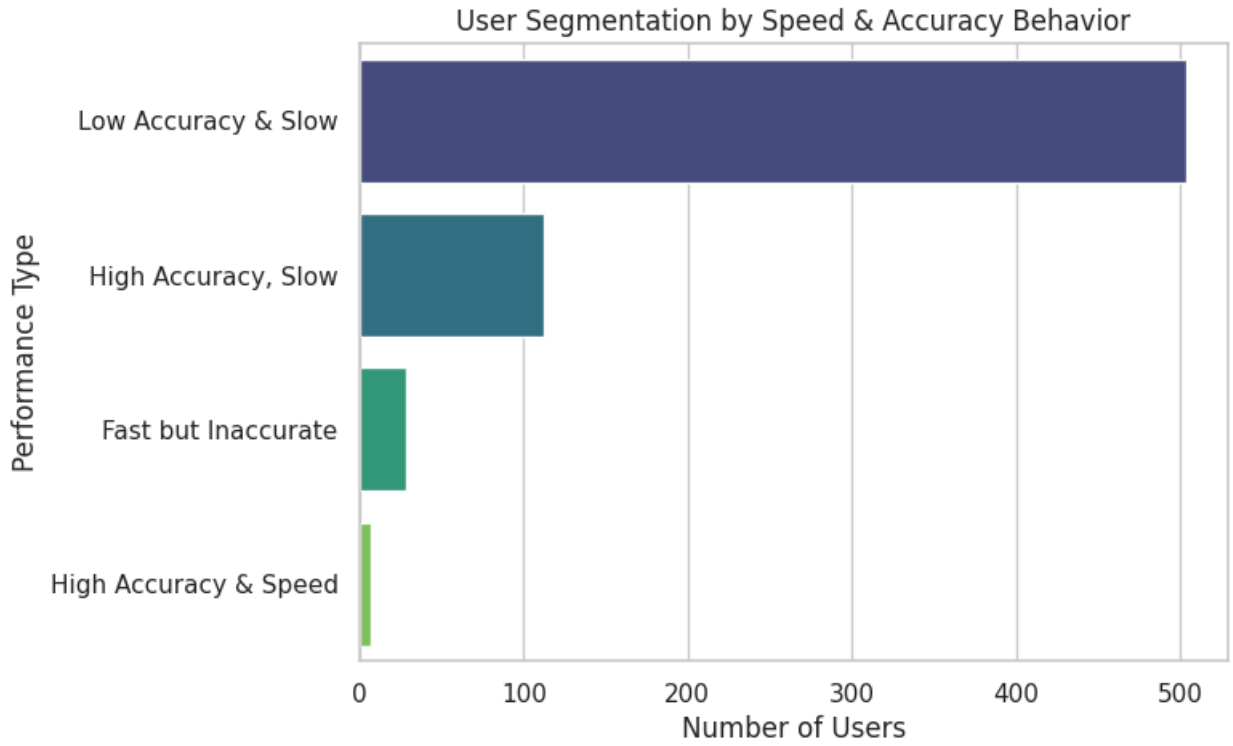
We defined A:S as accuracy score multiplied by speed score.
Below we can see the distribution of accuracy, speed and acc-speed ratio. 1 being the best score for both accuracy and speed. Which would mean someone had 50/50 correct answers and finished under a minute out of 60 minutes. To get an acc-speed ratio of 1 is nearly impossible.



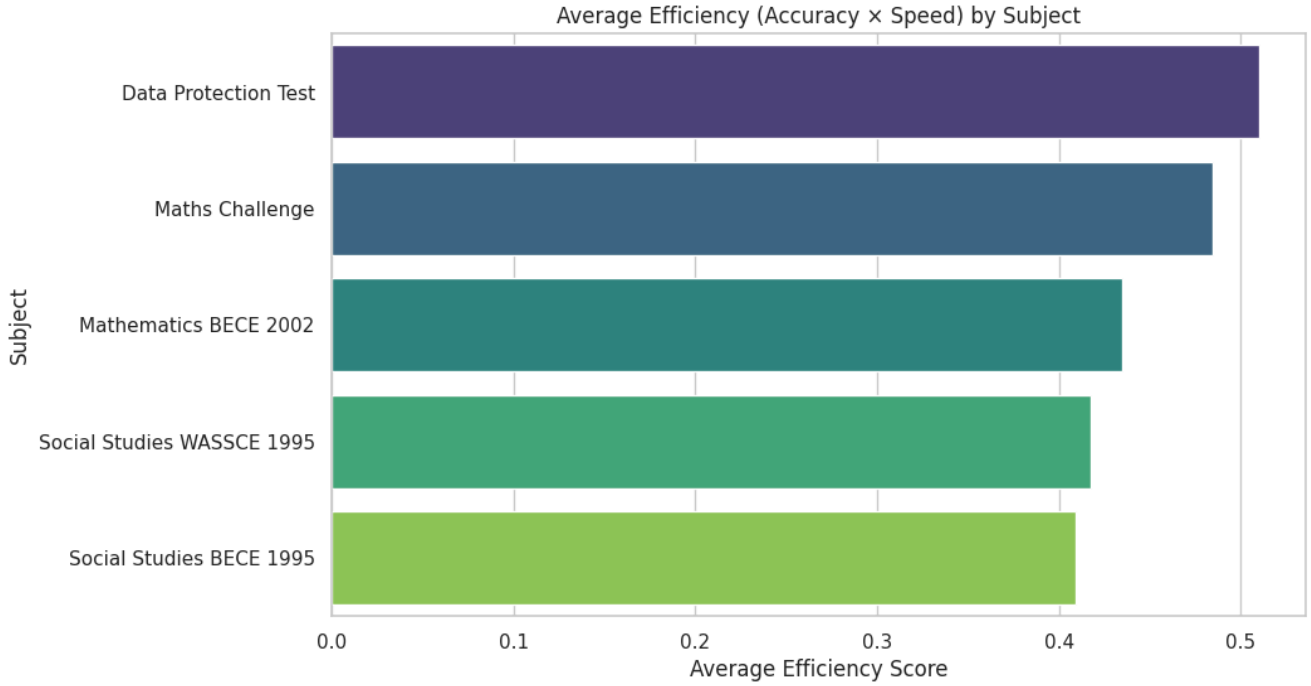
We further looked at user performance segmentation as seen in the plot below.



A more easier interpretation of user segmentation based on speed and accuracy below.

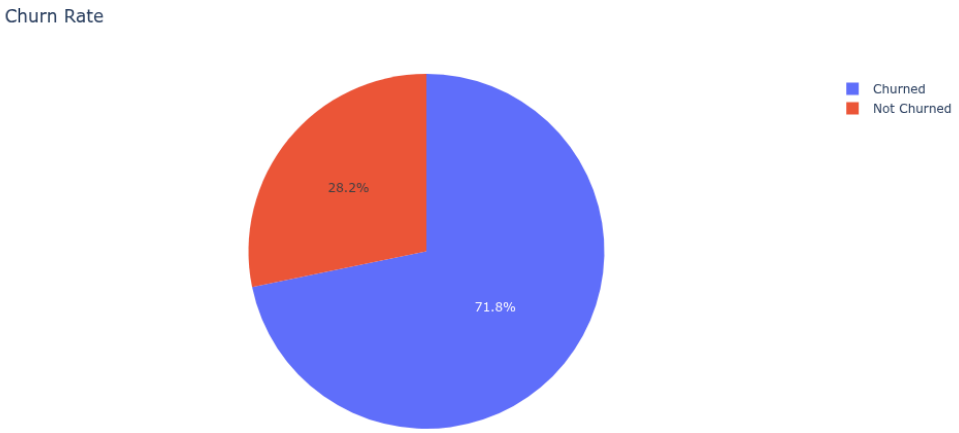


An analysis of courses/ subjects was also conducted based on the average accuracy to speed ratio on tests.

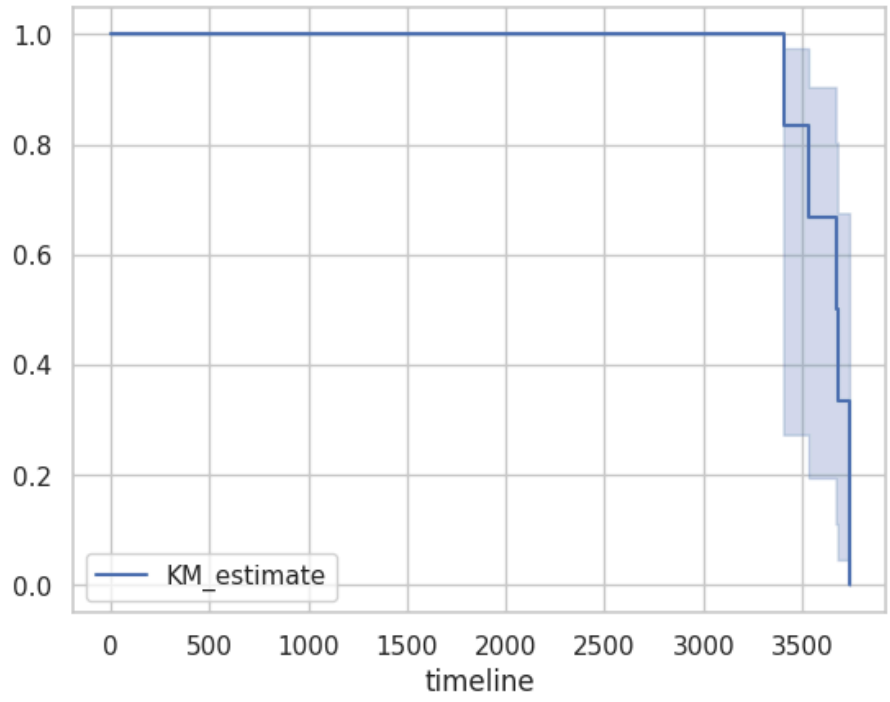


Churn and Risk (Initial Look):

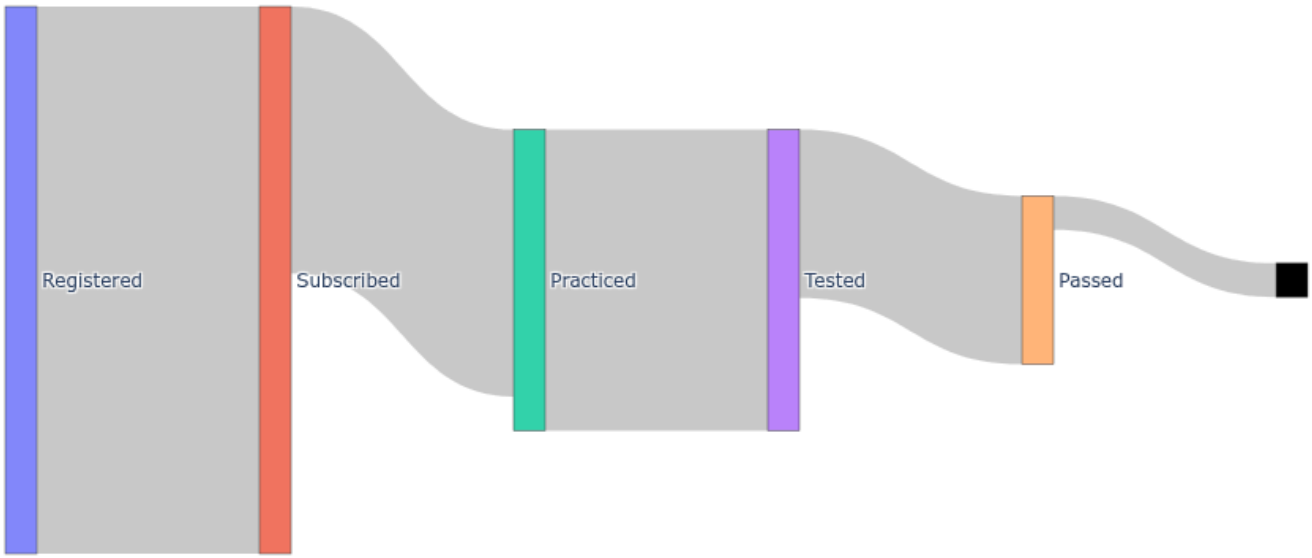
A basic churn rate based on 90 days of inactivity (using days_since_last_seen) was computed.
~72% of engaged users dropped off within 90 days



We could further explore where the churn is happening by institute, region etc to narrow it down. For the 28.2% that are still active (not churned) we conducted a survival curve, which gives a timeline to see how long it takes for them to churn (leave the platform). It would take about 3000 days on average.



Funneling user journey from account creation (registered) -> Subscriptions -> practice -> taking tests -> to pass results..

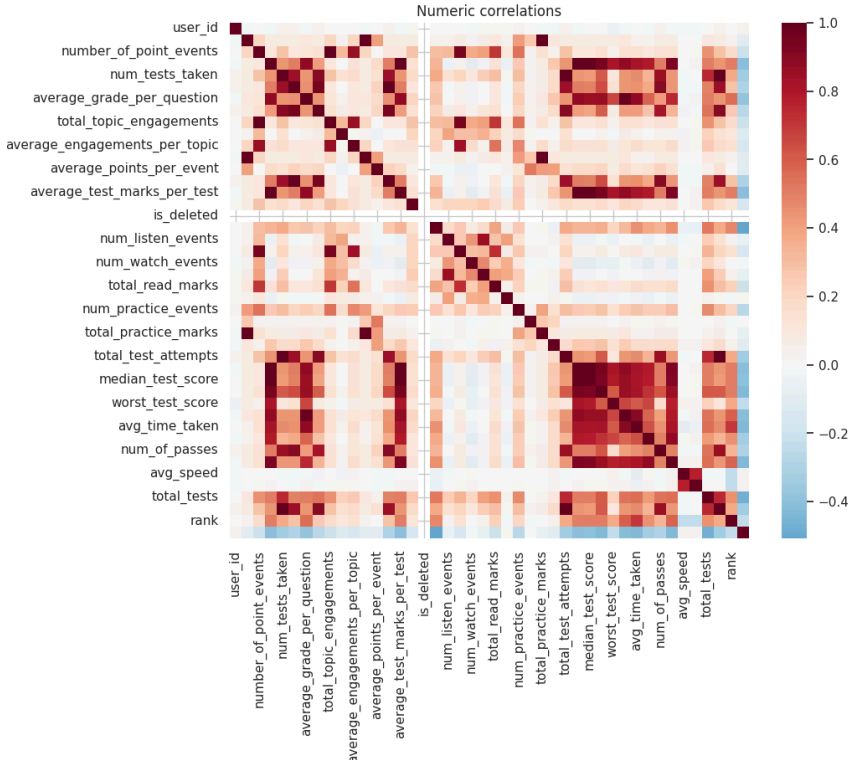


From the figure above we can observe the following:

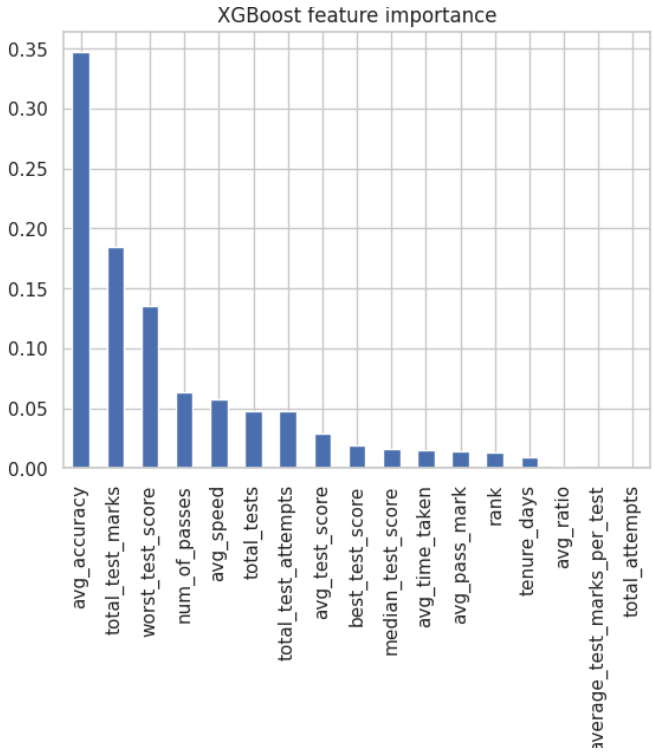
- an overall low conversation rate of ~6% from account creation to a user passing a test. That's from 2000 engaged users to 131 passes.
- a 48% drop in the number of users who subscribe after registrations.
- **12% increase in number of practice attempts.**
- 55% drop in users actually taking a test.
- 20% drop from test to pass.

Correlations and Feature Importance:

A correlation heatmap for numeric features was generated to identify initial relationships between different metrics.



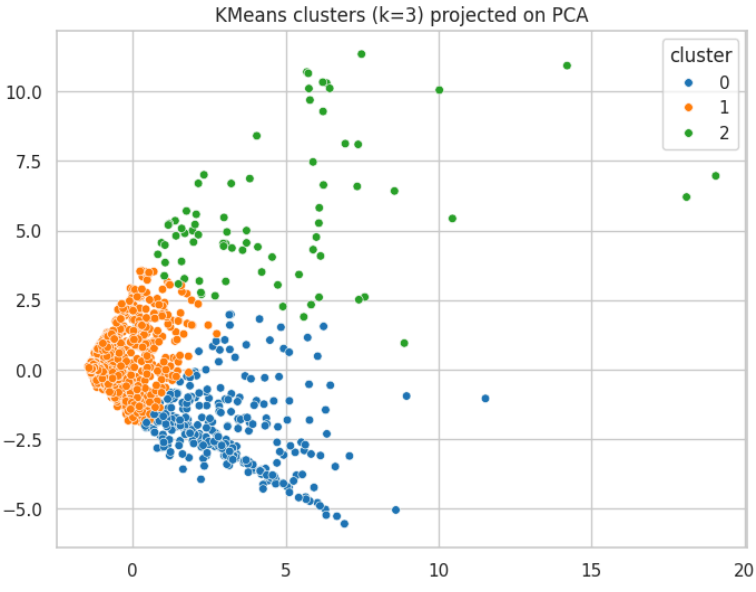
A quick XGBoost baseline was attempted to provide a preliminary look at potential feature importance for predicting a test outcome.



AUROC = 0.99

The **Area Under the ROC Curve (AUROC)** measures how well our model distinguishes between **students who pass vs. fail**.

- **Range:** 0.5 = random guessing, 1.0 = perfect discrimination.
- **0.91** means your model can correctly rank a random pair (one passing, one failing) **91% of the time**.
- That's excellent for an early-stage pilot — it shows your **feature engineering and data quality are strong**.



Clustering:

Clustering was performed on selected engagement features to identify distinct user segments based on their behavior.

We can see 3 separate clusters :

1. Highly engaging (blue); they spend more time on the app, read more, take more tests, and practice more. Average tenure is 1000+ days
2. Low / No activity (orange): hardly spend time on the app with average tenure being 14 days.
3. Mildly Active (green): casual users, less reading but more watching and listening content. Average tenure is 100+ days

Actionable Insights & Next Steps:

Prioritize Subscription Promotion: The strong correlation between the number of subscribed classes and retention highlights the importance of encouraging users to subscribe early. Business teams should explore incentives or product changes to drive initial subscriptions.

Investigate Low Engagement/Missing Data: The large number of users with no recorded activity in the test environment needs investigation. Understanding if this is a data issue, a test environment limitation, or reflects a real-world drop-off point is crucial. For production, ensure robust activity tracking.

Deep Dive into '0 Class' Users: A significant portion of the user base falls into the '0 Classes' category and shows low retention and high rates of missing activity data. Further analysis is needed to understand why they are not subscribing or engaging. Targeted interventions or onboarding flows could be designed for this segment.

Leverage Practice Data: The analysis of practice events and scores provides valuable signals for exam readiness. Further explore the relationship between practice behavior (frequency, scores, improvement) and test outcomes.

Address Low Pass Rate Topics/Tests: Identify tests or topics with consistently low pass rates. This could indicate challenging content that requires review, improved resources, or different teaching approaches.

Refine Churn Definition and Prediction: With real production data, refine the churn definition based on business goals and build predictive models to identify users at risk of churning.

Develop Early Warning Signals: Based on EDA findings and clustering, define simple heuristics (e.g., low points earned within the first month, no practice attempts after a certain period) to identify at-risk users for early intervention or nudges.

Utilize the Cleaned Data: The aggregated engaged user_profile_clean dataset provides a solid base for building predictive models and conducting deeper segmented analysis.

Recommendations

- From test Data it shows that lots of users use the platform more on tests and practice
- From a content perspective, reading seems to be the most active. Further exploration on what content is usually read and runtime will help in recommendations.
- Data Standardization: Have drop down entry options of city and country and other necessary columns to avoid case sensitivity.
- Inactive accounts removal.Define churn period and remove accounts that are not active.
- Funnel conversion rates (registered → practiced → tested → passed) with recommendations to address top drop-offs.
- Early-warning rules (heuristics) can be implemented before model ready (e.g., users with <2 practices in 14 days + tenure >90 days → nudge).
- Save user_profile_clean as the basis for future aggregated feature engineering and the 4000-student pilot.
- Early engagement recommendations upon sign up in the form of general and fun quizzes like ‘cities around the world’ or who won the cup in …? Or did you know?
- Collect data on gender and age for clustering and feature analysis. This could be done after initial signup to gain more points.
- User categorization based on educational level jhs/shs/tertiary/professional or our own custom methodology such a junior/senior/major/ pro etc
- More EDA to be performed with focus areas such as content consumption, practice vs test results and grading etc. (Mixpanel data can give more insights on content consumption.)