DSCI 599: Data Science for Business, Economics, and Society
Midterm Report

Trends in Online Courses: A Data-Driven Analysis

Yuheng Chen, Haoyue Xu, Jingyue Zhang

Team Members

Yuheng Chen, Haoyue Xu, Jingyue Zhang

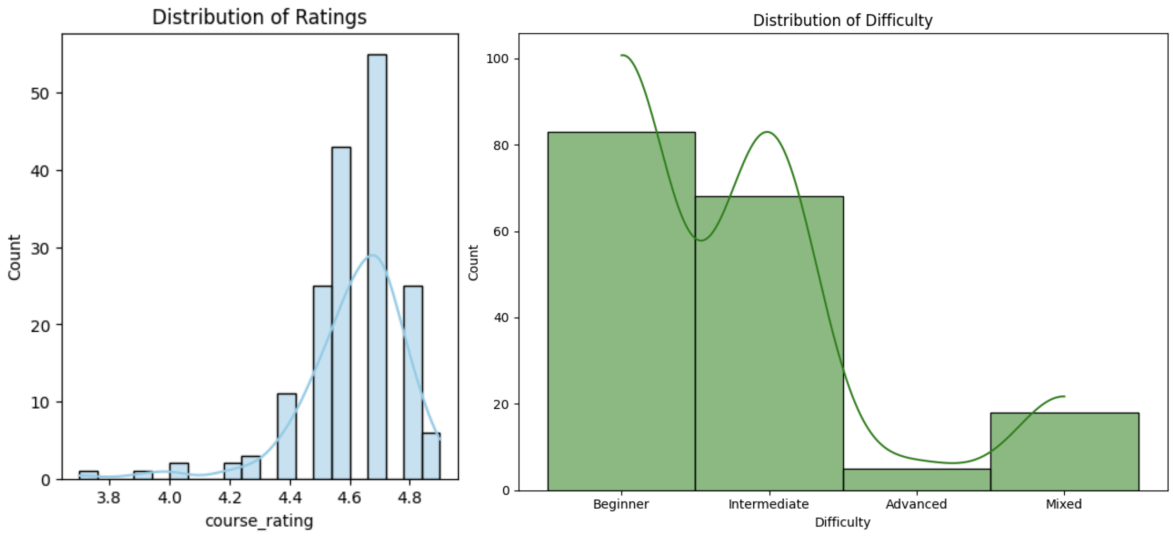
Working Progress

Our team has made significant progress in analyzing the Coursera datasets from 2021, 2023, and 2024. We have completed the following tasks:

Coursera Dataset 2021 Analysis

We filtered the dataset to focus on tech skill-related courses and examined the course ratings, difficulty levels, and enrollment numbers. We identified the top 10 technology-related courses and explored the correlation between course ratings and enrollment.

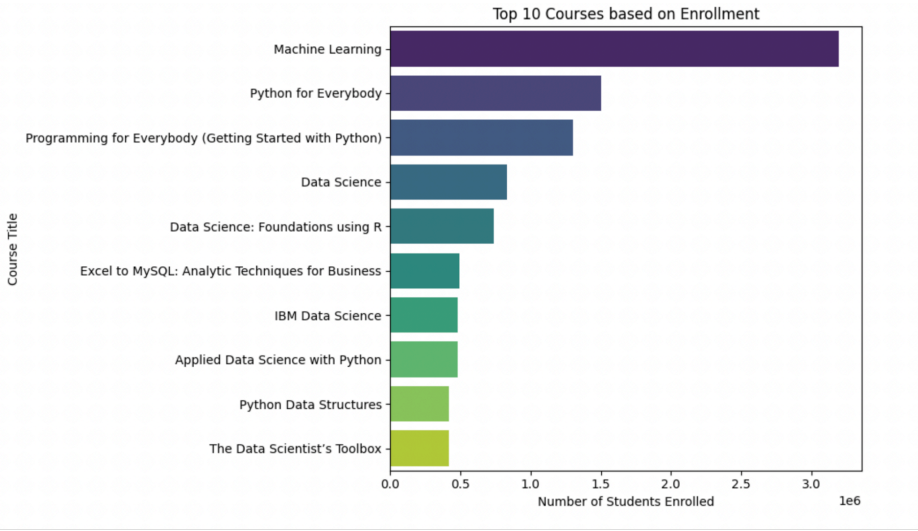
The dataset for Coursera courses from 2021 contains 892 entries. After filtering by tech skill-related keywords, it includes 174 entries. On average, these courses receive high ratings, with a mean rating of approximately 4.62. Most courses (75%) are rated below 4.7, indicating overall learner satisfaction.



Notably, beginner-level courses are prevalent, comprising 83 out of the 174 entries, suggesting a preference for accessible learning materials. The presence of fewer advanced classes, less than 10, may indicate that students with previous experience are less likely to choose Coursera for advanced topics.

Additionally, the mean enrollment per course is around 151,085 students, highlighting Coursera's appeal to a wide audience seeking educational enrichment in tech skills. In this bar

chart below, we have the top 10 technology-related courses from the Coursera 2021 dataset, ranked by enrollment numbers. The course with the highest enrollment appears to be Machine Learning. Other courses on the list focus on data science, Python, and applied data science . It's quite a clear indication of the popularity of machine learning and programming, especially with Python, among online learners. "Machine Learning" has significantly more enrollments than the others, crossing the 3 million mark.



A weak positive correlation (0.171) exists between course ratings and enrollment numbers, indicating that highly rated courses tend to attract larger audiences.

Coursera Dataset 2023 Analysis

We investigated the distribution of course ratings, difficulty levels, and certificate types. We found that most courses are highly rated and designed for beginners. We also discovered that the number of course enrollments and ratings are not related, but the number of enrollments and reviews are positively correlated. Additionally, we analyzed the distribution of certificates across different course durations and generated a word cloud to highlight the emphasis on technology and business education.

Coursera's analysis highlights its successful appeal to learners, primarily through high course ratings, a focus on beginner-friendly content, diverse certification options, and an emphasis on technology and business education. Course ratings reveal a concentration towards the higher end of the scale, with most ratings between 4.7 and 4.8, indicating widespread student satisfaction and a reflection of the platform's quality.

The analysis of course difficulty levels shows Coursera's strategy to cater mainly to beginners, with fewer advanced courses available. This choice aligns with the lower satisfaction ratings

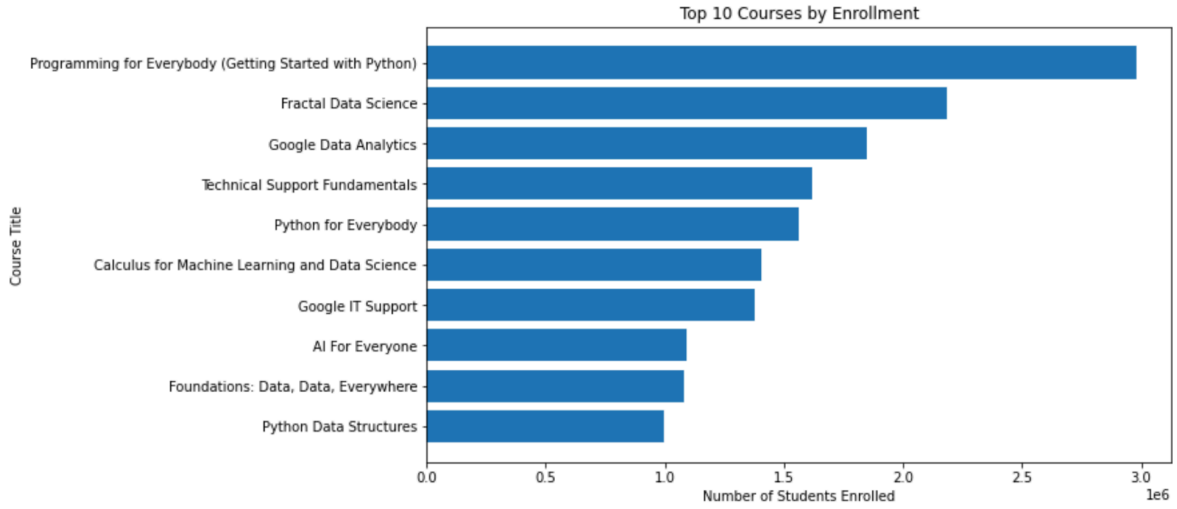
observed in more challenging courses, suggesting a nuanced relationship between course difficulty and student satisfaction.

In terms of certificate types, 'Course' certificates dominate, followed by 'Specialization', 'Professional Certificate', and 'Guided Projects'. This distribution points to a preference for specific skills or knowledge areas, with a significant portion of offerings designed for concise learning experiences. The majority of certificates are awarded for courses lasting 1 - 3 months and 1 - 4 weeks, catering to a demand for short-term, focused educational engagements.

Enrollment and review analysis uncovers that while course ratings and enrollments do not correlate, there is a positive correlation between the number of enrollments and reviews. This suggests popular courses generate more student engagement through feedback.

A significant portion of Coursera's content focuses on technology and business, with terms like "Data Science," "Machine Learning," "AI," "Python," and "Google" standing out. Among these, Machine Learning courses are particularly popular, representing 34.9% of enrollments, underscoring the high demand and interest in this field.

The top 10 courses by enrollment reflect this interest, with titles likely focusing on foundational aspects of Machine Learning, Data Science, and related technologies. These courses attract the most learners, evidencing the platform's alignment with current professional and educational trends.

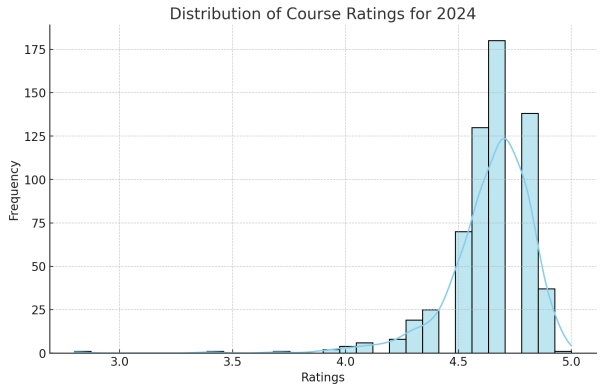


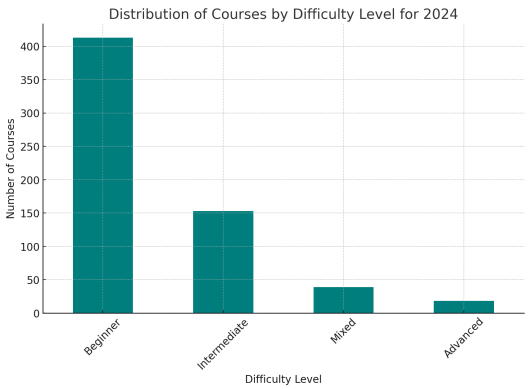
Coursera Dataset 2024 Analysis

We performed similar analyses on the 2024 dataset, examining the distribution of course ratings, difficulty levels, and certificate types. We found that the trends observed in 2023 continued in 2024, with beginner-level courses being the most common and having the highest average ratings. We also explored the correlation between course ratings and the number of

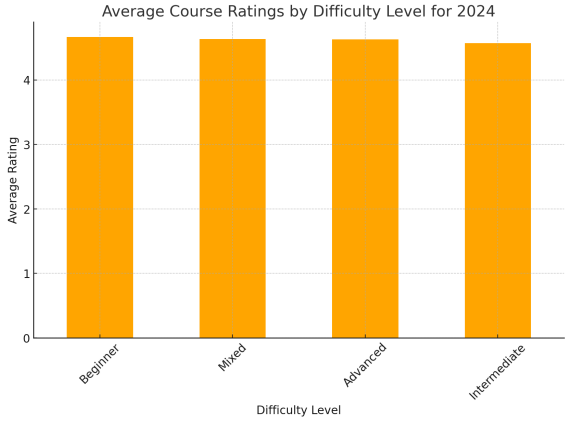
reviews, and analyzed the distribution of certificate types across different course durations. Lastly, we investigated the prominence of Machine Learning courses and identified the top 10 ML courses by enrollment.

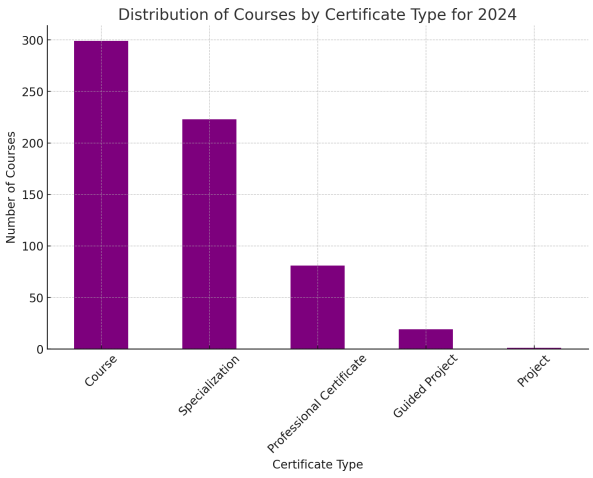
The distribution of course ratings for the 2024 dataset is shown in the following histogram. Similar to the 2023 analysis, it appears that the ratings are concentrated towards the higher end of the scale.





The distribution of courses by difficulty level for 2024 reveals that most courses are designed for beginners, with 413 courses categorized under this level. Intermediate courses follow with 153, mixed difficulty courses are at 39, and there are 18 advanced courses. This pattern aligns with the 2023 analysis, indicating that Coursera continues to predominantly cater to individuals who are new to the subject matter or looking to learn foundational skills.





The analysis of course ratings by difficulty level for 2024 shows that beginner-level courses have the highest average rating at approximately 4.67. Courses with a mixed difficulty level follow closely with an average rating of around 4.63, and advanced courses have an average rating of about 4.63 as well. Intermediate courses have the lowest average rating, at approximately 4.57. This trend suggests that, similar to the 2023 insights, the ratings of

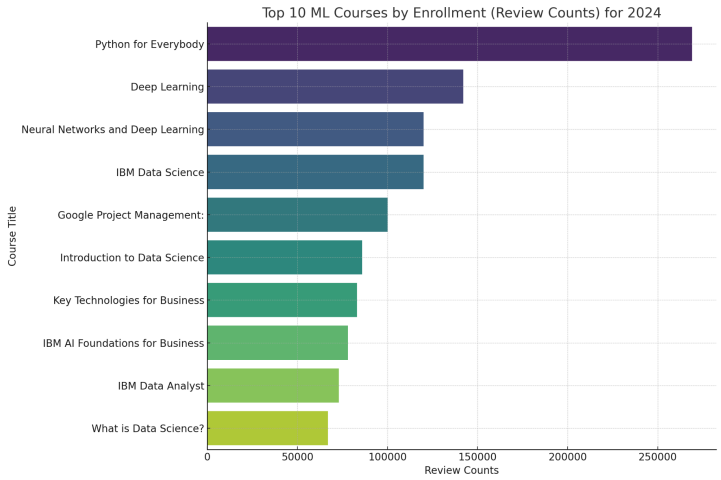
intermediate and advanced courses are generally lower than those for mixed and beginner levels, albeit the differences are slight.

In 2024, Coursera offers a diverse range of 623 certificate types, with 299 individual Courses being the most common, followed by 223 Specializations that provide in-depth subject exploration. There are 81 Professional Certificates aimed at career development, 19 Guided Projects targeting niche markets or emerging areas, and 1 Project, which might be an anomaly or a new category. This variety caters to various learning needs and objectives.

Course ratings and reviews have a slight positive correlation (0.164), suggesting higher-rated courses tend to have more reviews, but other factors also influence review count.

In 2024, shorter courses (1-4 weeks, 1-3 months) mostly offer individual "Course" certificates (298 total). Longer courses (3-6 months) emphasize "Specialization" certificates (196), followed by "Professional Certificate" courses (80). The "Other" category, likely for <2-hour courses, has 19 "Guided Project" and 1 "Project" certificates. This suggests a focus on "Specialization" for in-depth learning, "Course" for shorter durations, and "Guided Projects" for brief experiences.

In 2024, Machine Learning (ML) courses account for a substantial 37.15% of Coursera's student enrollment, based on review counts. The top 10 ML courses by enrollment span foundational to advanced topics.



Challenges

1. A significant challenge lies in ensuring the completeness and accuracy of the data collected over the years. With a vast and diverse array of courses, ratings, and student feedback, maintaining a high standard of data quality is crucial. Inaccuracies or missing data points can skew analysis and lead to misinformed decisions regarding course development and platform improvements.

2. Coursera's global platform means feedback and course descriptions come in a variety of languages. Analyzing this data requires tools that can understand different grammatical structures, idioms, and cultural nuances.
3. The difficulty in accessing comprehensive data on traditional university courses. Due to limitations in data availability and the often proprietary nature of educational content and outcomes from universities, we pivoted our focus towards online course data.

Updates on the Project Plan

Based on the three datasets provided, we have identified several promising avenues for future analysis to gain insights into the landscape of online education on Coursera. Our updated project plan includes the following key initiatives:

Comprehensive Trend Analysis
We will conduct an in-depth trend analysis to investigate the evolution of course offerings over time, with a particular focus on the emergence and growth of courses teaching specific skills. This analysis will help us understand how Coursera's course catalog has adapted to meet the changing needs of learners and the job market.

Student Engagement and Satisfaction
We will examine the relationship between course ratings, difficulty levels, and student enrollment numbers to identify the factors that drive student engagement and satisfaction. This information will be valuable for course designers and instructors to optimize their courses and improve learner outcomes.

Skills Gap Analysis
We will analyze the skills taught on Coursera with those in high demand in the job market to identify areas where course offerings could be strategically adapted to better prepare learners for the workforce.

Comparative Dataset Analysis
We will conduct comparative analyses between the different datasets to reveal shifts in focus or trends that might not be apparent when looking at a single dataset in isolation. This approach will provide a more comprehensive understanding of the online education landscape on Coursera.

Course Characteristics and Popularity
We will explore the relationship between course characteristics, such as duration and difficulty, and their popularity or ratings. This information will be valuable for course designers and instructors to optimize their courses for learner engagement and satisfaction.

By implementing these initiatives, we aim to gain actionable insights that will shape the future of learning on Coursera, benefiting the platform, its partners, and the millions of learners worldwide who rely on it for skill acquisition and career advancement.