Coursera Dataset: Exploratory Data Analysis

Presented By: Khristian Novoa

What is this project about?

In this project, I worked with a Coursera dataset that includes course titles, ratings, enrollment counts, difficulty levels, and certificate types.

My goal was to clean the data, explore trends, and use visualizations to better understand which courses and organizations are most impactful.

I used Python, specifically Pandas for data cleaning and transformation, and Matplotlib and Seaborn for creating all he charts you'll see in the upcoming slides.

Preparing the Data

Cleaning Steps:

- Removed unnecessary columns (e.g., Unnamed: 0)
- Converted enrollment values like "5.3k" and "1.2M" to numeric
- Filled missing values in course_students_enrolled with 0
- Standardized text formats (e.g., title casing for organization names)

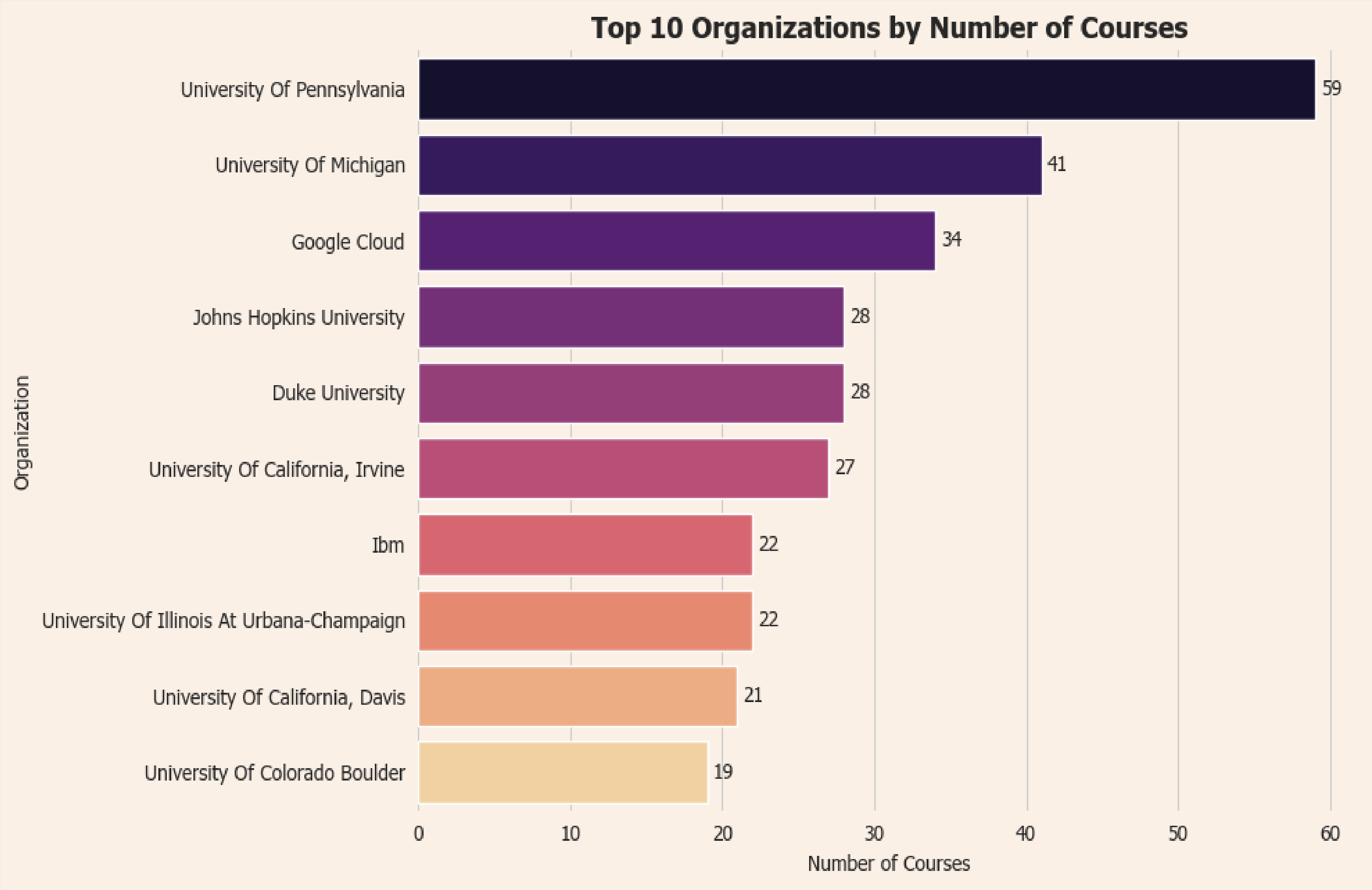
Data Checks:

- Used .info(), .isna().sum(), and .describe() to validate changes
- Verified data types and confirmed no NaNs in key columns

Result:

- A clean, structured dataset ready for analysis and visualization

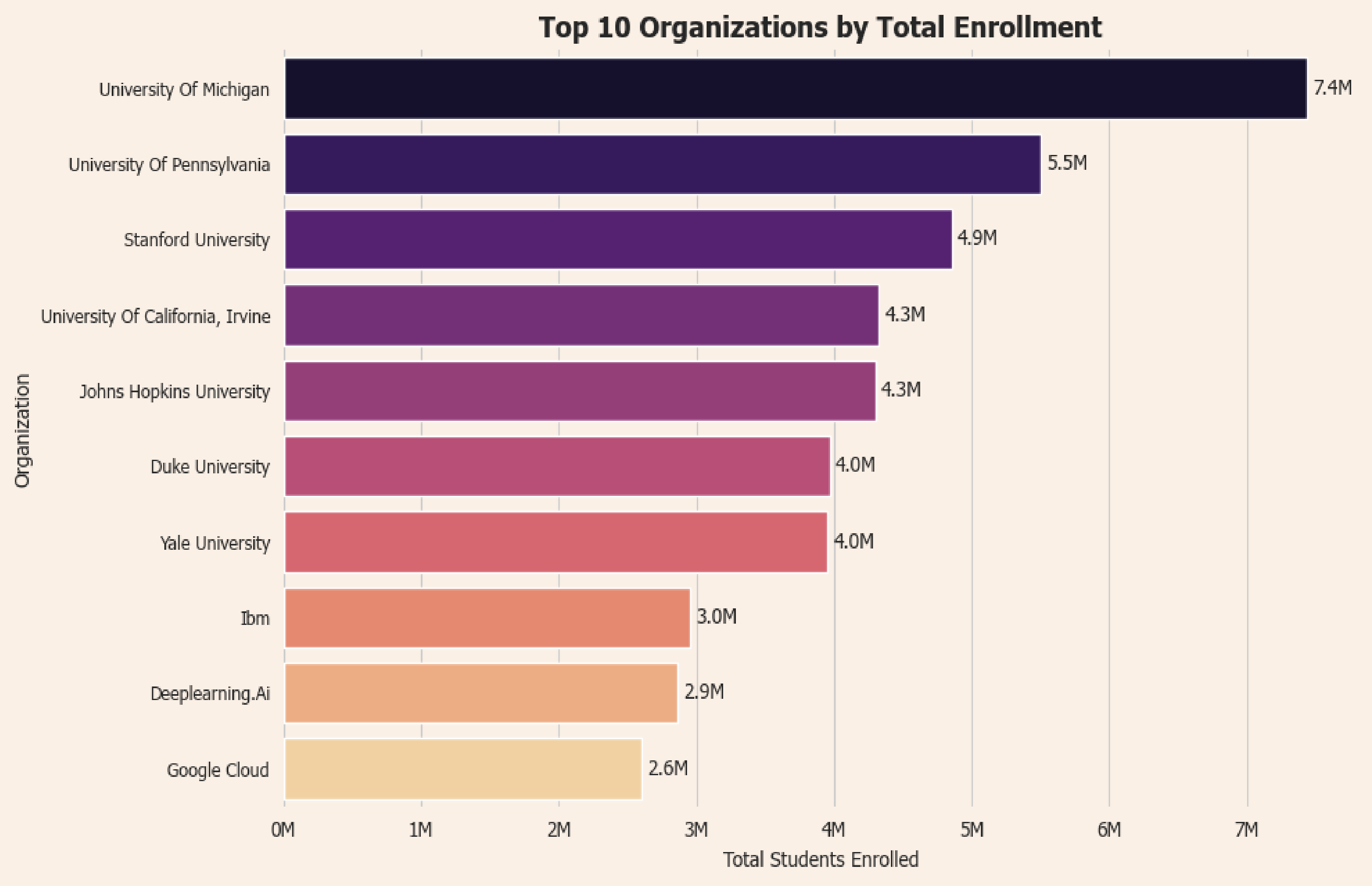
Which Organizations Offer the most Courses?



- **Top Organizations:**
- University of Pennsylvania
- University of Michigan
- Google Cloud
- Stanford University

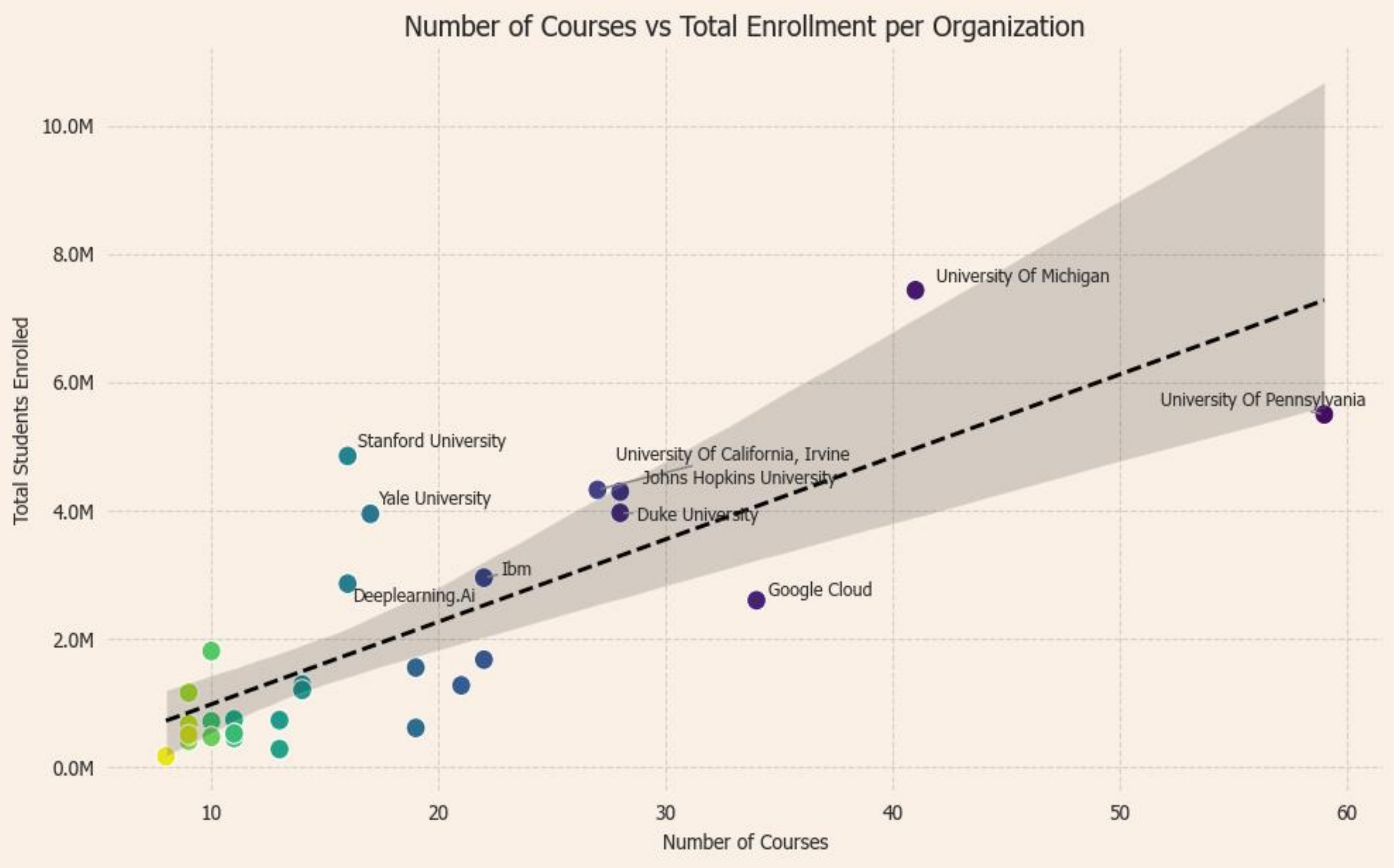
- **Key Insight:**
- Top contributors are mostly **universities**, with a few **tech companies**
- Indicates strong academic partnerships and professional development content

Which Organization is most Popular?



- Universities like **University of Michigan**, **University of Pennsylvania**, and **Stanford** have the highest total enrollments
- These organizations demonstrate massive reach across their course offerings
- Suggests strong brand recognition and high student interest

Do More Courses Lead to More Students?

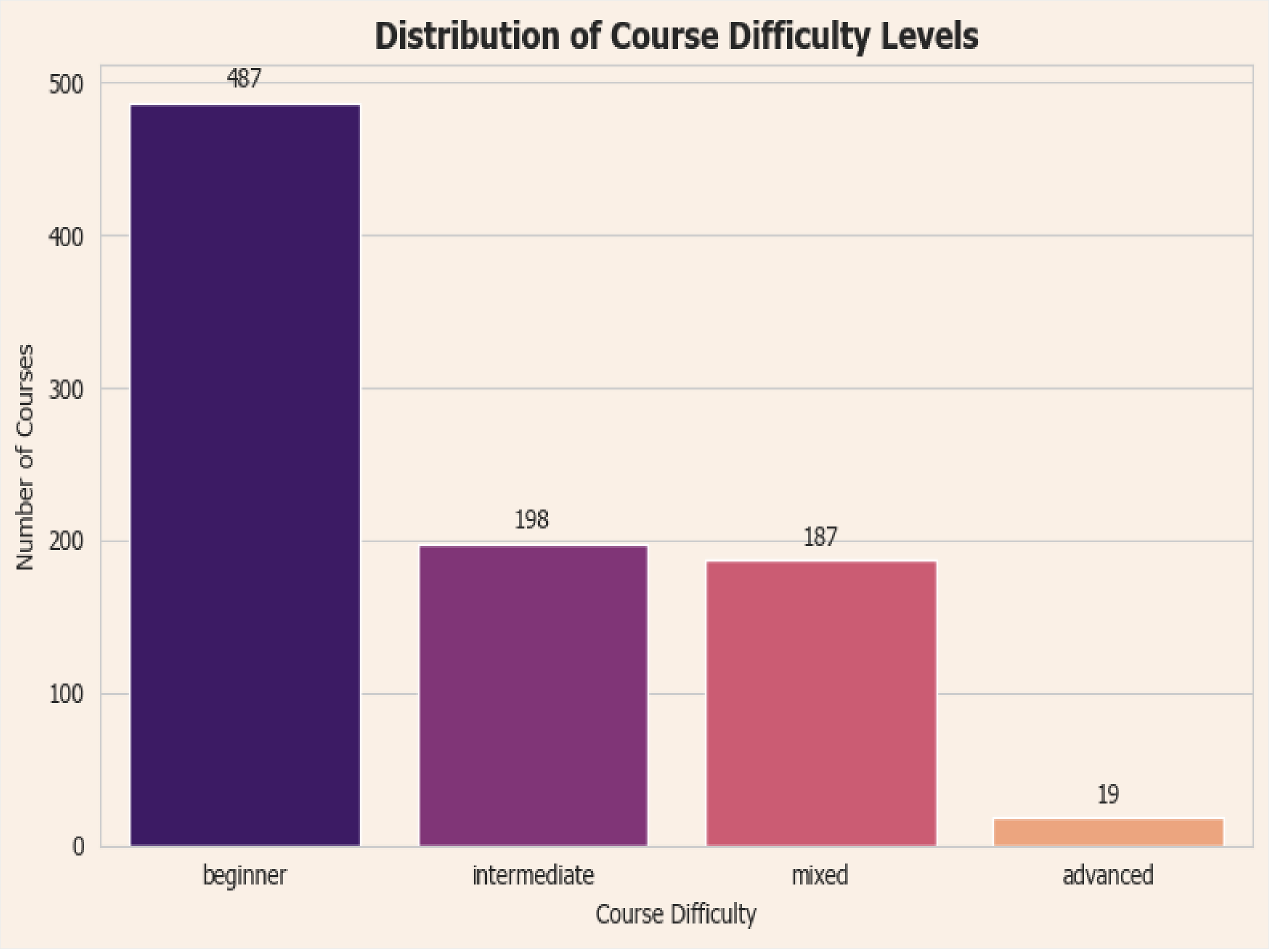


- The trend line shows a **positive correlation**, but with high variance
- Outliers like **Stanford** and **Yale** show **high reach with fewer courses**
- Organizations above the line are **overperforming**, those below may be under-leveraging their catalog

Key Takeaways

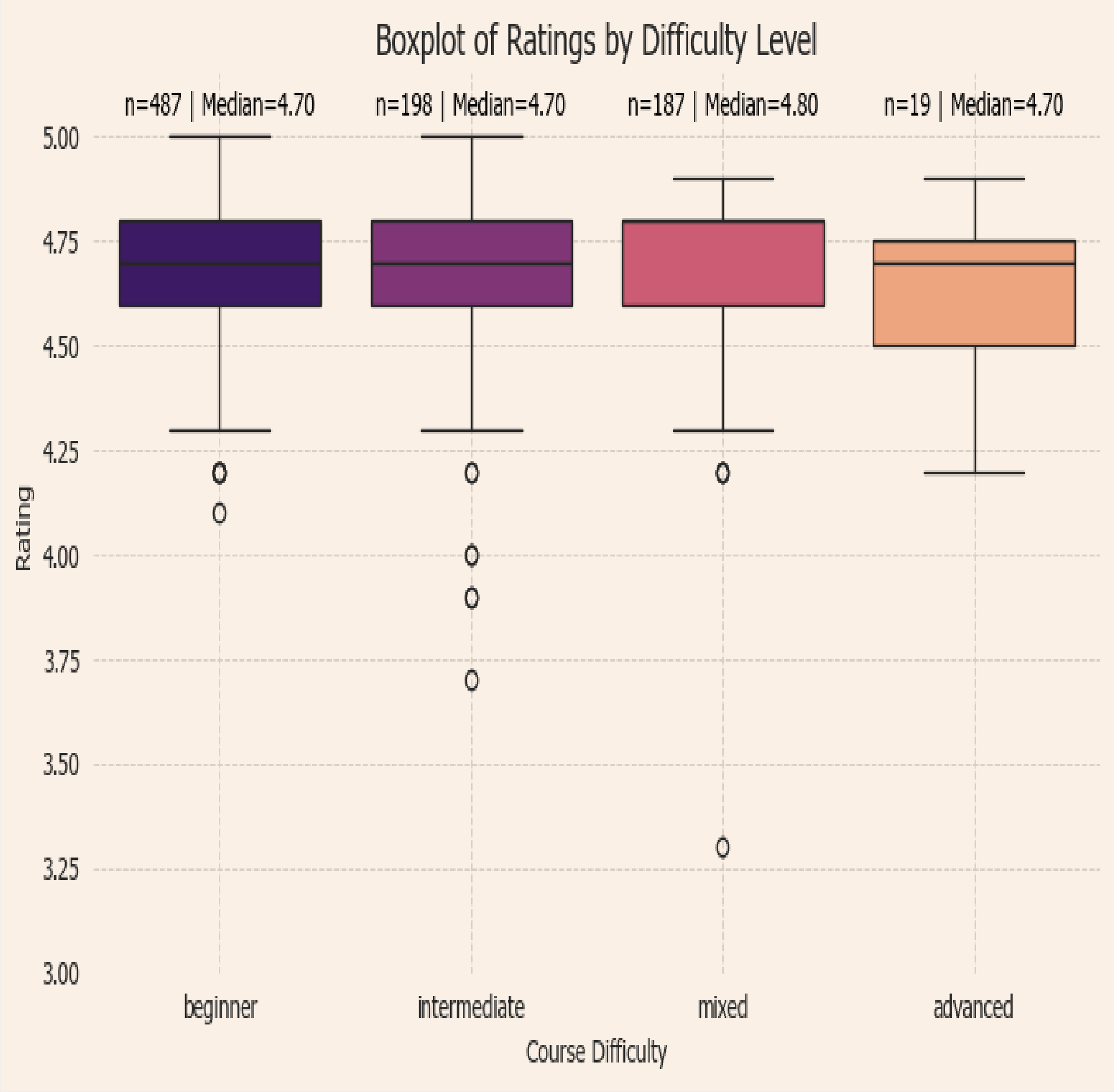
- **University of Pennsylvania** offers the most courses (59), followed by **University of Michigan** and **Google Cloud**
- **University of Michigan** leads in total student enrollment (~7.4M)
- **Stanford and Yale** achieve high enrollment with fewer courses — suggesting strong brand or course quality
- **Google Cloud** has many courses but lower total enrollment, highlighting that quantity doesn’t guarantee impact
- Organizations with **focused, high-quality offerings** tend to reach more learners per course

How are courses Distributed by Difficulty?



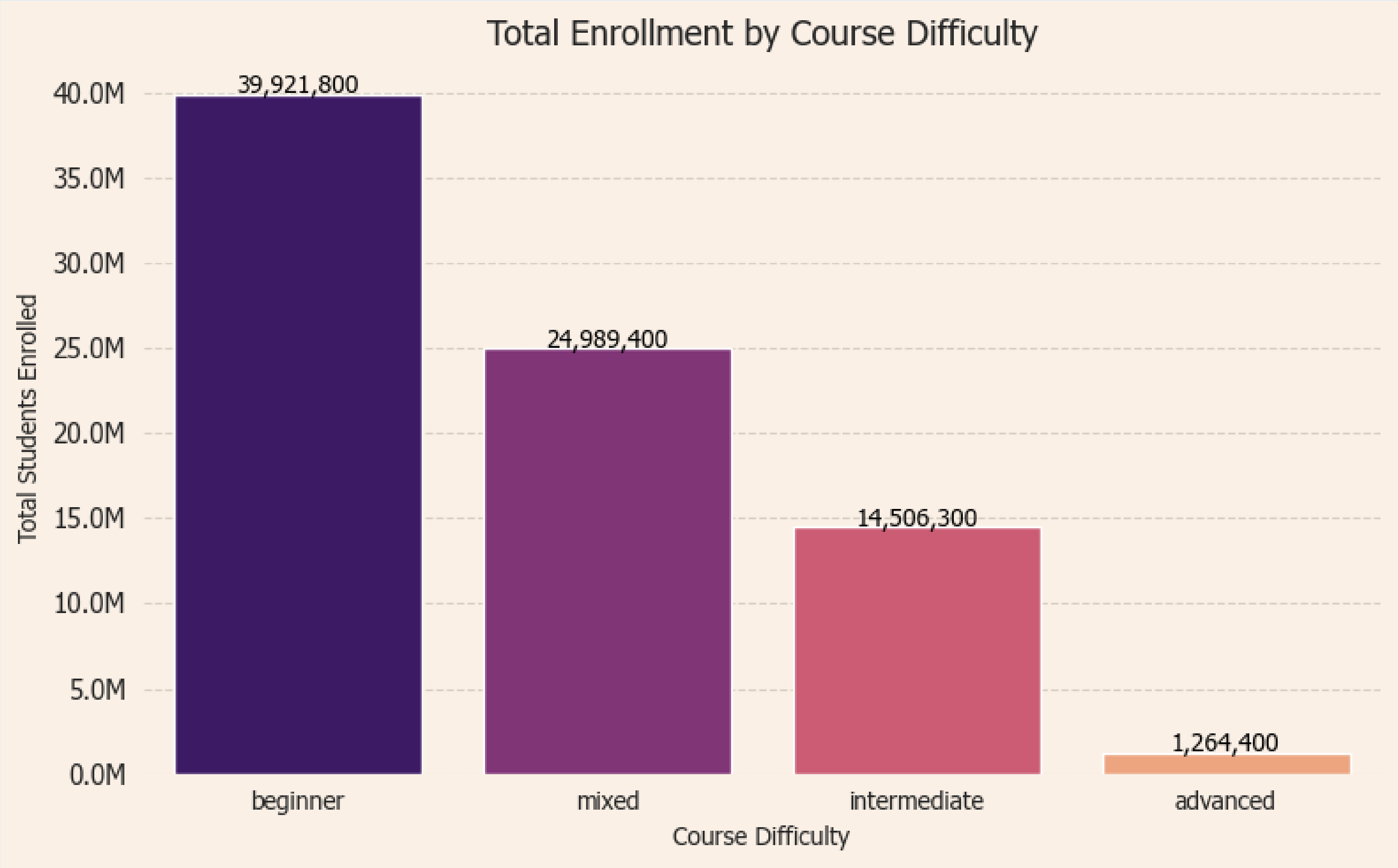
- Most Coursera courses are **beginner**
- Intermediate and Mixed levels make up a smaller portion
- Shows Coursera's emphasis on accessibility and entry-level learning

How Do Ratings Vary by Difficulty?



- All difficulty levels show **consistently high ratings**
- **Beginner, Intermediate, and Mixed** courses have a **median rating of 4.7**
- **Advanced courses are rated highest**, with a **median of 4.8**
- Suggests that learners are **satisfied regardless of difficulty level**

What Courses are Students Taking?



- **Beginner-level courses dominate**, reaching nearly **40 million students**
- **Mixed-difficulty courses** also perform well, with over **24 million enrollments**
- **Intermediate-level courses** trail behind at around **14.5 million**
- **Advanced courses** serve a small but specific group (~1.2 million students)
- Suggests Coursera’s largest impact is at the **entry-level and foundational skill tiers**

Key Takeaways

- **Beginner-level courses dominate** Coursera — both in number and total enrollment
- **Course quality is consistently high** across all difficulty levels
- **Advanced courses**, while few, are highly rated and serve a niche audience
- No strong link between **number of courses and total student engagement** — quality and content matter more than quantity
- Organizations can maximize impact by focusing on **student-aligned content** over just volume

Improvements and Future Work

- **Track Student Progression Across Difficulty Levels**
  - Analyze if students who take beginner courses go on to enroll in intermediate or advanced courses. This could help measure course effectiveness and student engagement over time.
- **Identify Multi-Course Learners**
  - Investigate how many courses individual students take. Understanding course stacking behaviors can highlight loyalty, satisfaction, or interest in specific subjects or platforms.
- **Analyze Trends Year Over Year**
  - Incorporate time-based analysis to see how course popularity, ratings, and enrollments change annually. This could uncover platform growth patterns, shifting student interests, or seasonal trends.
- **Retention & Drop-off Points (if data becomes available)**
  - It would be valuable to see where learners drop off or what keeps them coming back — this can shape course design or platform strategy.