**Duolingo's User Insights and Personalized Recap Dashboard**

**Problem Statement**

[Duolingo](https://www.duolingo.com/) wants to launch new features to make learning more fun and engaging by analysing how users learn and progress on the platform in order to **boost overall platform engagement and learning success rate.**

One of these features is **"Week in Review"**, which will show key details about each user and compare them to trends across the platform. It’s similar to [Spotify’s](https://medium.com/@kacperjohn/spotify-wrapped-2023-a-year-in-review-fef9cccfdc47) or [YouTube’s](https://blog.tunemymusic.com/youtube-music-recap-learn-how-to-get-youtube-music-stats/) "Year in Review."

As a data analyst, your tasks are:

1. **Analyse overall user data (EDA) to f**ind trends and behavior patterns and suggest improvements to make the platform better and explain why.
2. **Create a personalized recap for the most engaged user (t**he user who is most active on the platform) by comparing their performance to overall platform trends and sharing their overall metrics.

The goal is to uncover insights that will:

* Help Duolingo improve its strategies to keep users engaged and learning.
* Provide users with a fun, personalised recap to make learning more rewarding and engaging.

**Focus on the following:**

1. Understanding of how to analyze user engagement and learning success on the Duolingo platform.
2. Clear articulation of the purpose behind analyzing the dataset.
3. Identification of relevant metrics from the dataset to guide the analysis.
4. Definition of specific and testable hypotheses.
5. Inspection of dataset structure and data types.
6. Identification and handling of missing or duplicate values.
7. Detection and management of outliers in key metrics.
8. Cleaning and preprocessing the dataset for analysis.
9. Normalization of values for better comparisons.
10. Grouping and aggregating data to identify trends.
11. Conducting exploratory data analysis (EDA) to uncover trends and patterns.
12. Testing hypotheses and evaluating relationships between key metrics.
13. Development of visualizations to effectively present findings.
14. Use of correct visuals to highlight trends, comparisons, and relationships.
15. Structured presentation of findings based on analysis and visualizations.
16. Data-supported insights aligned with the objectives of the analysis.

**Deliverables**

1. Jupyter Notebook
   * Scripts for data cleaning, processing, analysis and visualisation.
   * Outputs showing key findings from the analysis for both overall trends and personalized user data.
   * [Sample Jupyter Notebook](https://colab.research.google.com/drive/1lhganARhkmFNCSG8rT8Voy2nipsU1Zhj?usp=sharing) built on Google Colab (*Please look at all cells carefully)*
   * Use [Google Colab](https://colab.research.google.com/) to run the code and queries.
2. Tableau Dashboard
   * Interactive dashboard with visualisations of key metrics with proper filters and storyboard.
   * User-friendly layout to make insights accessible for stakeholders.
3. PDF document Sharing the insights
   * A structured report highlighting business context, key metrics and hypothesis.
   * Key findings from the overall EDA and documentation of trends, patterns, and insights supported by visualisations and data analysis.

**IMP NOTE:**

1. Name should not be present in any of the files shared i.e. Google Colab Jupyter notebook and PDF document.
2. All file names will be saves as "\*\*NL Duolingo.<File extension\*\*>"
3. Maximum limit of PDF document is 15 pages and all text should be above font size 12 and font style Times New Roman or Arial.

**Dataset**

Here is the dataset link - <https://drive.google.com/file/d/1Jhl4lWVCMsvuWzTVY7hYCE88AMv-f1-7/view?usp=drive_link>

Please use below code to import it to Google Colab.

# Step 1: Install gdown

!pip install gdown

# Step 2: Import the file from Google Drive

import gdown

# File ID and URL

file\_id = "1Jhl4lWVCMsvuWzTVY7hYCE88AMv-f1-7"

url = f"https://drive.google.com/uc?id={file\_id}"

# Output file name

output = "duolingo\_data.csv"

# Step 3: Download the file

gdown.download(url, output, quiet=False)

# Step 4: Load the file into a DataFrame

import pandas as pd

data = pd.read\_csv(output)

**Here's an explanation of the data columns:**

Imagine we’re analyzing a Duolingo lesson where a student is learning the word “cat” (in Spanish: “gato”). Here's what each column would represent:

1. p\_recall (Proportion of Recall Accuracy)
   * Definition: The proportion of exercises in this lesson where the word (or lexeme) was correctly recalled by the student.
   * Example: If the student encountered "gato" in 5 exercises and correctly recalled it in 4, the p\_recall value would be 0.8 (4/5).
2. timestamp (Time of the Lesson)
   * Definition: The timestamp indicating when the current lesson or practice took place.
3. delta (Time Gap)
   * Definition: The time (in seconds) since the last lesson or practice where this specific word (lexeme) was encountered.
   * Example: If the student last practiced "gato" 2 days ago (48 hours = 172800 seconds), the delta value would be 172800.
4. user\_id (Student ID)
   * Definition: An anonymized ID representing the student who completed the lesson or practice.
5. learning\_language (Language Being Learned)
   * Definition: The target language that the student is learning.
   * Example: If the student is learning Spanish, this column would show "es" (language code for Spanish).
6. ui\_language (User Interface Language)
   * Definition: The language of the app’s user interface, which is usually the student's native language.
   * Example: If the student’s app is in English, this column would show "en".
7. lexeme\_id (Lexeme Tag ID)
   * Definition: A system-generated unique ID for the word or lexeme being practiced.
   * Example: If “gato” is represented by the ID 123abc456, this column would show 123abc456.
8. lexeme\_string (Lexeme Tag)
   * Definition: A detailed grammar tag describing the lexeme (word), including its properties like tense, gender, and plurality.
   * Example: For “gato” (a masculine singular noun), the tag might look like "gato", where:
     + means noun
     + means masculine
     + means singular
   * More Details on lexeme tags is shared [here](https://docs.google.com/spreadsheets/d/11PboM69gIuK2V51l_pK-WEwjkPTpGFuLGCWVggzG7l8/edit?gid=0#gid=0)
9. history\_seen (Times Seen Before)
   * Definition: The total number of times the student has encountered this word (lexeme) in lessons or practice sessions before this one.
   * Example: If the student has seen “gato” in 10 exercises before, this column would show 10.
10. history\_correct (Times Correct Before)
    * Definition: The total number of times the student correctly recalled this word (lexeme) in previous lessons or practice sessions.
    * Example: If the student has correctly recalled “gato” in 7 out of 10 past exercises, this column would show 7.
11. session\_seen (Times the Word/Lexeme Was Seen in the Current Session)
    * Definition: This column indicates how many times the student encountered the specific word or lexeme during the current lesson or practice session.
    * Example: Imagine the student is learning "gato" in their current lesson. During this session:
      + The student encounters "gato" in 3 different exercises (e.g., a multiple-choice question, a fill-in-the-blank exercise, and a sentence translation).
      + The value in the session\_seen column would be 3, representing the total encounters with "gato" in this session.
12. session\_correct (Times the Word/Lexeme Was Correctly Recalled in the Current Session)
    * Definition: This column indicates how many times the student correctly recalled or answered the specific word or lexeme during the current lesson or practice session.
    * Example: Continuing with "gato":
      + The student was correct in 2 out of the 3 exercises where "gato" appeared during the current session.
      + The value in the session\_correct column would be 2, representing the total number of correct responses for "gato" in this session.