User Segmentation Analysis for Duolingo

-GurSimran K Sujlana

Introduction :

This analysis was conducted with the objective of identifying distinct user segments or personas among Duolingo users. These segments are intended to inform product development and marketing strategies to enhance user engagement and satisfaction.

Data Preparation :

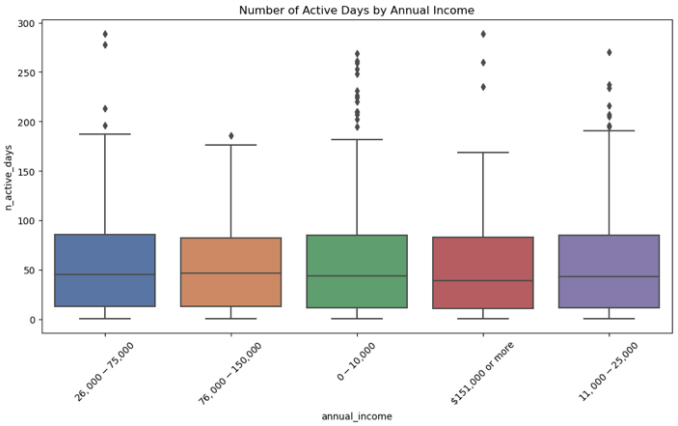
Data Cleaning and Pre-processing :

The dataset comprised a series of survey responses and usage data from Duolingo users. Initial steps involved cleaning the data, which included addressing missing values and ensuring data consistency.

Visualization :

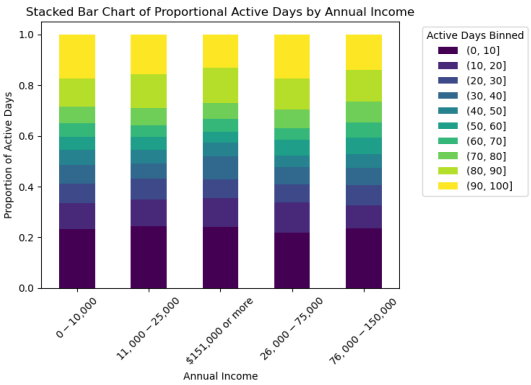
Highest Crown Count by Age Group :

Younger users (18-34) show a higher range and slightly higher median in crown counts, indicating greater engagement or proficiency, with substantial variability within each age group.



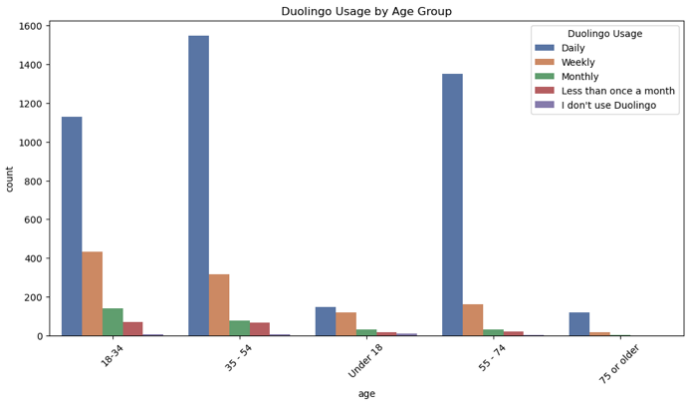
Number of Active Days by Annual Income :

Higher income groups tend to show a slightly increased median in active days, suggesting more consistent engagement, yet the variability is similar across all income levels.



Duolingo Usage by Age Group :

The '18-34' age group is the most active daily user demographic, while daily usage declines with age, and the 'Under 18' group shows notably less daily engagement.



Handling Missing Values with MICE :

Missing values were imputed using the Multivariate Imputation by Chained Equations (MICE) method, ensuring that the richness of the dataset was preserved and that subsequent analyses had a robust foundation.

Text Feature Encoding with BERT :

Text responses were encoded using the BERT (Bidirectional Encoder Representations from Transformers) model, converting open-ended text data into meaningful numerical representations that could be utilized in the analysis.

Feature Engineering and Standardization :

Numerical features were standardized to have a mean of zero and a standard deviation of one. This standardization is crucial when performing Principal Component Analysis (PCA) to ensure that all features contribute equally to the result.

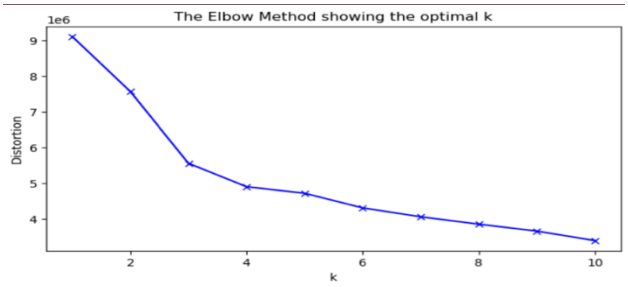
Dimensionality Reduction with PCA :

PCA was applied to reduce the complexity of the data while retaining the variation present across the different users. This step transformed the high-dimensional data into a lower-dimensional space where clusters could be more easily identified.

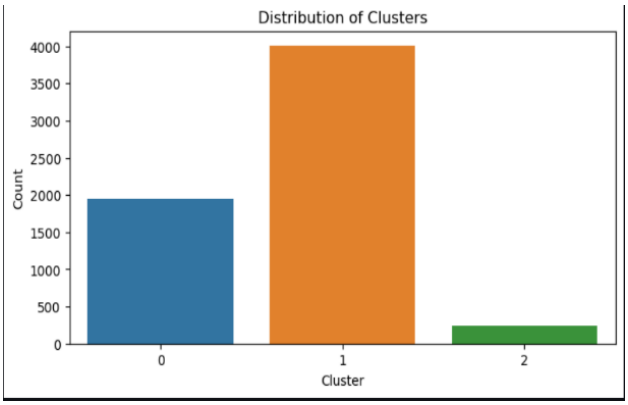
Clustering with KMeans :

The KMeans clustering algorithm was employed to identify clusters within the PCA-transformed space. The optimal number of clusters was determined using the Elbow Method, which indicated three distinct clusters.

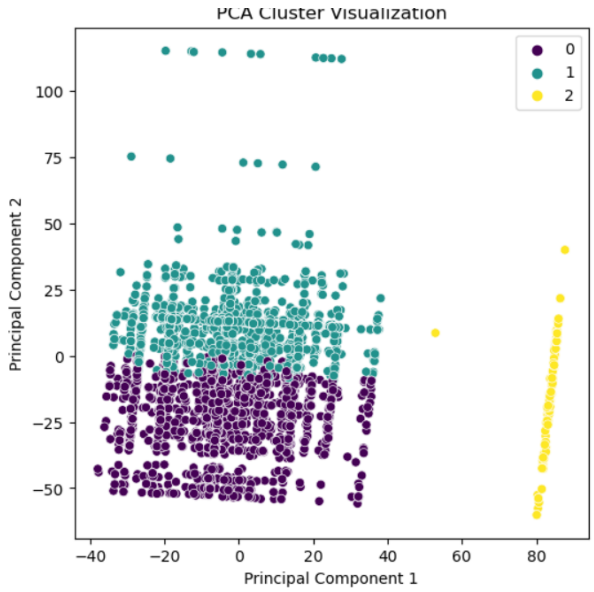
Determining Number of clusters:



Distribution of Clusters :



Visualization of Clusters :



Insights and Inferences

Cluster 0: The Engaged Learners

Cluster 0 was characterized by users with lower overall engagement but consistent content preferences. These users might benefit from features that encourage regular engagement, such as daily challenges or streaks.

Cluster 1: The Exploratory Users

Cluster 1 consisted of users with varied content preferences and learning behaviors. This cluster suggested a need for personalized content recommendations and diverse learning paths.

Cluster 2: The Committed Achievers

Cluster 2 represented highly engaged users with specific goals or motivations for using the service. This segment could be targeted with advanced features, tracking, and analytics to support their learning objectives.

Strategic Recommendations :

Product Development

For Cluster 0, develop features that support habit formation and consistent usage.

For Cluster 1, enhance the recommendation engine to offer a personalized experience.

For Cluster 2, introduce goal-oriented features and advanced tracking.

Marketing Strategies

For Cluster 0, create campaigns that highlight the benefits of regular practice.

For Cluster 1, showcase the variety of content available to cater to diverse interests.

For Cluster 2, use success stories and advanced feature highlights to attract users with specific learning goals.

Cross-Cluster Strategies

- Implement community-building initiatives to leverage the knowledge of highly engaged users to mentor others.
- Establish feedback mechanisms to refine the user experience based on direct user input.
- Conduct surveys to capture changing user needs and preferences.

Conclusion

The analysis successfully identified three user segments with distinct characteristics and behaviours. The strategic recommendations provided aim to cater to the unique needs of each cluster, with the goal of enhancing overall user engagement and satisfaction on the Duolingo platform.