Technical Appendix- From Free to Premium: Understanding Spotify's User Segments

Dream Stream Team
BANA 205 Team 18B: Shivani Vallamdas, Christie Shin, Vishal Srivastava, Gema Zhu, Shuai Zhao

I. Data Description

This *Spotify User Behavior Dataset* used in this project is sourced from Kaggle, a public data-sharing platform. The dataset is openly available, and no special permissions are required for academic analysis. It consists of anonymous survey-based responses that capture users' listening habits, demographic characteristics, and engagement behaviors. The dataset includes 520 observations and 20 variables, of which we focus on the following eight variables for our clustering analysis:

1) *Age* - Age group of user?
2) *Gender* - Gender of user?
3) *spotify_usage_period* - How long have you been using Spotify?
4) *spotify_subscription_plan* - Which Spotify subscription plan do you currently have?
5) *premium_sub_willingness* - Are you willing to subscribe to the premium plan or keep your premium subscription?
6) *music_time_slot* - What is your favourite time slot to listen to music?
7) *music_Influencial_mood* - When it comes to listening to music, which of the following moods or situations most strongly influences your choice of music?
8) *music_context* - When do you listen to music more often?

II. Data Processing

Excel Data Cleaning & Manipulation (Using Python)

Format Correction: We addressed specific formatting inconsistencies caused by Excel auto-formatting to preserve demographic accuracy.

E.g. Age: '20-Dec' (Date Error) = '12-20' (Text String), '12-Jun' (Date Error) = '6-12' (Text String)

Missing Value Handling: We standardized non-informative text strings to ensure the clustering algorithm correctly interprets missing data.

E.g. Null Values: 'None' (String literal) = `NaN` (Numeric Null)

*We also changed *music_lis_frequency* to *music_context* for better clarity and understanding.

Dimensionality Transformation (Explode Method): We applied a split-and-explode transformation to multi-select variables to capture the full depth of user behavior (User-Context Moments).

Examples:

- Mood influencing music choice: "Relaxation, Sadness" (Single Cell) = Split into 2 Separate Rows
- Music context: "Workout, Travel" (Single Cell) = Split into 2 Separate Rows

Strategic Segmentation: We physically bifurcated the dataset to analyze distinct revenue streams separately.

E.g. Subscription Cohorts: Combined Dataset = Split into 'Free Users' and 'Premium Users' file**s**

Encoding Method

We categorized the variables into numerical form to ensure compatibility with KMeans, which requires numeric inputs. This method allows us to include all variables into the clustering process.

1) *Age*: '6-12', '12-20' = 1, '20-35' = 2, '35-60'= 3
2) *Gender*: Male = 0, Female = 1, Other = 2
3) *Spotify usage period*: 'Less than 6 months' = 0, '6 months to 1 year' = 1, '1 year to 2 years' = 2, 'More than 2 years' = 3
4) *Premium subscription plan*: Free = 0, Premium = 1
5) *Willingness to continue Premium*: No = 0, Yes = 1
6) *Listening time slot*: Morning=0, Afternoon=1, Night=2
7) *Mood influencing music choice*: "Relaxation and stress relief" = 0, "Uplifting and motivational" = 1, "Sadness or melancholy" = 2, "Social gatherings or parties" = 3
8) *Music context*: "Random" = 0, "Office hours" = 1, "Study Hours" = 2, "While Traveling" = 3, "Workout session" = 4, "Before bed" = 5, "Night time" = 6, "Social gatherings" = 7, "leisure time" = 8, "when cooking" = 9

III. Analytical Methods

Software and Tools

We used two main tools throughout the analysis: Python and Excel.

- **Python (pandas, scikit-learn, matplotlib, seaborn):** Python served as the core environment for all data processing and modeling. We used pandas for cleaning and transforming the dataset, scikit-learn to run KMeans clustering, silhouette score evaluation, and PCA, and matplotlib/seaborn to visualize both cluster patterns and dimensionality-reduction results. Python was chosen because it provides a flexible, reproducible workflow and integrates all machine-learning techniques required for this project.
- **Excel:** Excel was used during the early exploratory phase to manually inspect the variables, verify missing values, and review basic descriptive statistics. Its spreadsheet interface made it easier to quickly confirm data quality before transitioning to Python for more complex modeling.

Techniques Employed

Several analytical and machine learning techniques were applied to understand user behavior and identify meaningful segments within the dataset.

- **KMeans Clustering:** We used this technique to group users based on key behavioral and demographic variables because it performs well with numerical data and is effective for identifying distinct user segments. This approach allows us to uncover clusters that differ not only in who the users are, but also in how, when, and why they listen to music.
- **Silhouette Score Evaluation:** We implemented this score to determine the optimal number of clusters for our analysis because it measures cluster cohesion and separation, helping validate whether the clustering structure is meaningful.
- **Principal Component Analysis (PCA):** This was applied to reduce the dataset's dimensionality to make it easier to interpret the clustering results. With this two-dimensional visualization, it helped confirm whether the KMeans clusters formed meaningful, specific groupings.

Model Building

Selected Variables and Exclusions

For our clustering analysis, we included all variables directly related to music listening behavior, demographics, and subscription preference. Podcast-related variables were intentionally excluded because our research objective focuses specifically on music consumption patterns, rather than general Spotify usage. We felt that including those variables would introduce unnecessary noise into the clustering process and reduce clarity on the segments we aimed to identify.

Model choice: KMeans vs. Logistic Regression

We selected KMeans clustering for this analysis because our goal was to discover natural groupings of Spotify users based on their music behaviors, demographics, and subscription characteristics. Since there is no predefined outcome variable to predict, a supervised method like logistic regression would not be the best option. In contrast, an unsupervised method like KMeans enables the data to reveal distinct clusters that emerge organically from listening habits. This makes KMeans a more suitable approach for our study.

Optimal Cluster Choice: K=3

Our Silhouette Score analysis indicated that while a higher number of clusters (K=7) yielded a marginally higher statistical score (0.27 vs 0.23), we selected K=3 as the optimal model for this business case.

We prioritized interpretability and actionability over raw statistical variance for these key reasons:

1. Avoidance of Micro-Segmentation At higher K values (e.g., K=7)

The model began splitting users into redundant micro-groups based on minor demographic traits rather than major behavioral differences. For example, distinguishing between "Morning Pop Listeners" and "Afternoon Pop Listeners" offers little strategic value, as the retention strategy for both remains identical.

2. Distinct Behavioral Profiles With K=3

The clusters crystallized into three non-overlapping, behaviorally distinct profiles that map directly to the user lifecycle:

Cluster 0: The low-engagement/casual user.
Cluster 1: The high-potential/context-driven user.
Cluster 2: The established loyalist.
This separation allows for clear, contradictory marketing strategies (e.g., "Educate" vs. "Upsell" vs. "Retain") without ambiguity.

3. Resource Efficiency From a deployment perspective

K=3 allows the marketing team to focus their budget on three high-impact campaigns. Managing 7 distinct segments would dilute marketing resources and complicate campaign execution for diminishing returns in conversion rates.

Conclusion: K=3 was selected because it offers the maximum business insight with the minimum complexity, providing the most robust foundation for a scalable segmentation strategy.

References

Ajayakumar, M. (2023, July 6). *Spotify User Behavior Dataset*. Kaggle. https://www.kaggle.com/datasets/meeraajayakumar/spotify-user-behavior-dataset