Background

Spotify has gained significant market share in the music streaming service industry, generating a profit and beating analyst expectations in Q2 2024. Spotify offers a free membership with ads or users can pay a monthly subscription to eliminate ads and receive access to all features on their platform. The company sought to build a business model that attracts users to its platform and away from the massive amounts of digital piracy that occurred during the 2000s. Consumers settled and chose the platform which offers access to music and other content for a small monthly price over pirating music and storing it locally on their devices. The result is a positive relationship between creators, record labels, Spotify, and listeners.

Motivation

Spotify Web API allows developers to retrieve metadata from its content, and users can request access to listening data through their profile settings. This gives users the ability to analyze their music trends over time to better understand themselves. An individual's music taste grants insights about their personality, essentially creating a proxy psychological profile. Our analysis aims to evaluate my personal Spotify data to develop this proxy characterization and create product recommendations for a future marketplace. The marketplace serves as medium for creators to sell products (i.e., merchandise, widgets, etc.) to consumers through the Spotify application. Product recommendations in the marketplace are based on both the user's current music trends and the profile created about them.

The analysis will primarily focus on:

- **Which artists and genres are the most popular?**
  The bulk of an individual's profile will come from this information and the demographic data user's input to create their account (i.e., birthday, gender, address, and Facebook account, if applicable) and establish baseline generalizations about a person.
- **What audio features are the most prevalent?**
  The Spotify Web API includes the audio features for each song, such as acousticness, danceability, duration (milliseconds), energy, liveness, loudness, speechiness, tempo, instrumentalness, and valence. These features will be used to make further generalizations. For example, songs with high danceability, energy, and tempo might indicate the user's level of physical activity.
- **Can we identify how time impacts music trends?**
  Time data is used to understand a user's sensitivity to specific holidays (i.e., seasonality). Spikes in American-themed songs at the beginning of July might indicate a strong level of patriotism, so American-themed products have a higher probability of selling. Time information is also used at the daily level to make assumptions about a user's routine. If songs with high tempo, danceability, and loudness are primarily played during the evening, we can assume the user most likely works out at this time.
- **Can playlists confirm our generalizations?**
  A user's playlists provide valuable insights since they are created specifically by the user. Playlist information is used to make further and seeks to confirm generalizations from patterns about the user identified elsewhere.

Description of the Data

The initial data requested from Spotify is comprised of 2 .zip files containing data stored as JSON files. The JSON files include user, playlist, search query, and streaming history data. The data was requested from the profile settings in my Spotify account. The dataset spans from 2018-08-17 (the time my account was created) to 2024-09-01. After combining all the streaming data into one CSV file, our initial dataset contains 203,317 observations with 21 variables. The main variables from this dataset include timestamp, track name, track artist, track album, and a Spotify track ID.

Using our initial dataset, the Spotify Web API will be queried to extract the audio features for all songs and saved as a separate CSV file. Then, our audio features dataset will be joined to our initial dataset and subsequent views generated to perform our analysis. Since Spotify is the owner of both data sources, the data received is high quality and mostly clean. The only caveat is the two datetime variables. The **offline_timestamp** variable contains multiple zero UNIX values, so variable **ts** must be used to convert timestamps to the desired date format.

Proposed Analysis

Our Spotify analysis will utilize k-means clustering, time series analysis, random forest models (TBD), ARIMA models, and exploratory analysis/visualizations such as boxplots and histograms for exploratory analysis.

K-means clustering is used to help understand the underlying structure of the user's listening history. The method will analyze each song's audio features to group songs with similar characteristics into clusters. These clusters will help lay the foundation to start building the proxy psychological profile of a user. If I can get my friends and coworkers to volunteer their data, our analysis will also analyze listening patterns to cluster users. Clustering users will enable the creation of targeted marketing strategies for our theoretical Spotify marketplace.

Time series analysis is used to identify seasonality among genres in the user's listening patterns. Significant increases in genre song count around specific holidays indicates positivity; therefore, the product recommendations section of the Spotify marketplace would start to suggest products related to that holiday. Music creators can capitalize off this opportunity because recommendations are prioritized by the artist's popularity with the user, and the user can shop the artist's merchandise in a centralized location.

ARIMA models are used to confirm findings generated from the time series analysis. These models can identify popularity and seasonal trends for a song or genre over time to predict future listening trends. For example, a song/genre's popularity in the upcoming weeks can be forecasted to increase, remain constant, or decline based on the user's historical listening patterns. Successful forecasting creates opportunities to deploy specific marketing strategies to influence their purchasing behavior in the marketplace.

Exploratory analysis/visualizations are used to identify patterns through various aggregations and statistical summaries. Histograms are used to analyze the distributions of our audio features, and correlation plots to observe correlations between two features. Timestamps are analyzed to discover any consistent patterns between genres/clusters and the time of day to understand the user's listening patterns. Stated previously, songs with high loudness, energy, and tempo that are most frequently being played in the evenings might suggest the user routinely works out at night. Most of a user's proxy psychological profile is generated from various aggregations in the exploratory analysis and clustering, but due to current data being limited to one person and a theoretical marketplace, our models will only predict the likelihood of the user purchasing a holiday product.

References

"Spotify for Developers" Spotify, https://developer.spotify.com/
"Understanding My Data" Spotify, https://support.spotify.com/us/article/understanding-my-data/

Appendix

The raw data was requested from my Spotify profile settings.

- Includes observations from 2018-08-17 to 2024-09-01.

Spotify Web API: https://developer.spotify.com/documentation/web-api/

The data was received as two .zip folders containing various JSON files.
The streaming history JSON files were merged into one CSV file with 203,317 observations with 21 variables.
The variables are described as follows:

- **ts:** Date and time of when the stream ended in UTC format.
- **username:** Your Spotify username.
- **platform:** Platform used when streaming the track (e.g. Android OS, Google Chromecast).
- **ms_played:** For how many milliseconds the track was played.
- **conn_country:** Country code of the country where the stream was played.
- **ip_addr_decrypted:** IP address used when streaming the track.
- **user_agent_decrypted:** User agent used when streaming the track (e.g. a browser, like Mozilla Firefox, or Safari).
- **master_metadata_track_name:** Name of the track.
- **master_metadata_album_artist_name:** Name of the artist, band or podcast.
- **master_metadata_album_album_name:** Name of the album of the track.
- **spotify_track_uri:** A Spotify Track URI, that is identifying the unique music track.
- **episode_name:** Name of the episode of the podcast.
- **episode_show_name:** Name of the show of the podcast.
- **spotify_episode_uri:** A Spotify Episode URI, that is identifying the unique podcast episode.
- **reason_start:** Reason why the track started (e.g. previous track finished or you picked it from the playlist).
- **reason_end:** Reason why the track ended (e.g. the track finished playing or you hit the next button).
- **shuffle:** Whether shuffle mode was used when playing the track.
- **skipped:** Information whether the user skipped to the next song.
- **offline:** Information whether the track was played in offline mode.
- **offline_timestamp:** Timestamp of when offline mode was used, if it was used.
- **incognito_mode:** Information whether the track was played during a private session.