Spotify Songs' Genre Segmentation Report

This report outlines a data analysis project using the Spotify dataset to segment songs by genre, visualize audio features, cluster tracks, and build a recommendation system. The process uses Power BI for visualizations and Python for preprocessing and clustering, with results presented in a clear, concise dashboard.

1. Data Preprocessing

- **Dataset:** Spotify dataset (Spotify dataset.csv) with features like track popularity, danceability, energy, loudness, and playlist genres.
- **Steps:**
  - Loaded data and checked for missing values and duplicates.
  - Removed duplicates based on track_id and dropped rows with missing values.
  - Selected numerical features: danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence, tempo.
  - Normalized features using StandardScaler for clustering.
  - Encoded categorical features (playlist_genre, playlist_subgenre) as numbers.
  - Saved pre-processed data as spotify_preprocessed.csv.
- **Power BI:** Imported spotify_preprocessed.csv, verified data cleanliness in Query Editor.

2. Data Visualizations

Visualizations were created to explore feature distributions and relationships, using Power BI's native tools and Python-generated images.

- **Bar Chart:** Average track popularity by genre (pop and EDM often have higher popularity ~50–60).
- **Histograms:** Showed distributions of numerical features (e.g., danceability peaks at 0.6–0.8 for dance pop).
  - Python Visual scripts for each feature (e.g., energy, loudness) used seaborn.histplot with KDE curves.

- **Correlation Matrix:** Heatmap showed strong correlations (e.g., energy and loudness ~0.7).
- **Box Plot**: Compared danceability across genres (EDM median ~0.7, rock ~0.5).
- **Scatter Plot:** Tempo vs. danceability, colored by genre, showed EDM tracks cluster at high tempo and danceability.
- **Pie Chart**: Genre distribution (pop ~30–40% of tracks).
- **Power BI Setup:** Used native visuals (bar, scatter, pie), custom visuals (Box and Whisker, Correlation Plot), or imported Python images (e.g., correlation_matrix.png).

3. Clustering

- **Method:** K-Means clustering on normalized features to group similar tracks.
- **Steps:**
  - Used elbow method to select 4 clusters.
  - Applied K-Means, added cluster labels to dataset (spotify_clustered.csv).
  - Visualized clusters in Power BI (scatter plot: danceability vs. energy, colored by cluster).
- **Insights:** Clusters grouped tracks by audio profiles (e.g., high danceability/energy for dance pop).

4. Recommendation System

- **Method:** Content-based system using cosine similarity on normalized features.
- **Steps:**
  - Computed similarity matrix for tracks.
  - Built function to recommend top 5 similar tracks for a given track ID (e.g., “I Don’t Care” by Justin Bieber).
  - Saved recommendations as spotify_recommendations.csv.
- **Power BI:** Displayed recommendations in a table visual with a track_id slicer.

5. Power BI Dashboard

- **Overview Page:** Pie chart (genre distribution), bar chart (popularity by genre).
- **Feature Analysis Page:** Histograms, box plot (danceability by genre), scatter plot (tempo vs. danceability), correlation matrix.
- **Clustering Page:** Scatter plot (clusters by danceability/energy), bar chart (clusters by playlist).
- **Recommendations Page:** Table of recommended tracks with slicer.

6. Key Insights

Pop and EDM genres have higher popularity and danceability (~0.65–0.7).

Energy and loudness strongly correlate (~0.7), especially in EDM.

Clusters align with genres (e.g., high-energy tracks in EDM playlists).

Recommendations match tracks with similar audio profiles, often within the same genre.

7. Interesting Fact

Tracks in "Pop Remix" playlists have high danceability (~0.65) and energy (~0.85), ideal for upbeat playlists, but some popular tracks (e.g., "Higher Love" by Kygo) have lower danceability (~0.69) yet high valence (~0.40), evoking positive emotions.