UNDERSTANDING USER MUSIC PREFERENCES ON SPOTIFY

FINAL REPORT

IST-652 SCRIPTING FOR DATA ANALYSIS

GROUP 5

- KHUSHI SHETTY
- AYUSH SASEENDRAN
- MALVIKA DIWAN

INTRODUCTION

The music industry is rapidly evolving, and data analysis is playing a pivotal role in shaping user experiences and music recommendations. In this project, we aim to explore the potential of data analysis in the context of Spotify.

Our objective is to gain a comprehensive understanding of user music preferences on Spotify, exploring playlist creation, track listening patterns, and favored genres. By delving into these aspects, we aim to derive valuable insights into user behavior, offering a nuanced perspective on how users engage with and curate their musical experiences within the Spotify platform. This analysis will unveil patterns, trends, and emotional nuances, contributing to the enhancement of user experiences in the dynamic realm of digital music consumption.

DATA

The primary data which the project uses is a rich dataset sourced from Spotify's Web API. The extracted data was converted to JSON datasets which offer comprehensive information about tracks in various playlists. This data encompasses:

- **Track Details:** Names, artists, and album information.
- **Audio Features:** Quantitative measures like danceability, energy, valence, tempo, loudness, and speechiness.
- **Lyrics:** Extracted from an external lyrics API, providing a textual dimension to the audio data.

We have also used other data for the sentiment analysis of the top 6 songs of the user. Data is hosted on Amazon S3, showing cloud storage utilization for the positive and negative word lists.

- **Positive Words:** Fetched from https://intro-datascience.s3.us-east-2.amazonaws.com/positive-words.txt.
- **Negative Words:** Obtained from https://intro-datascience.s3.us-east-2.amazonaws.com/negative-words.txt.

Another text file was used for emotion analysis on two of the playlists:

- The NRC Emotion Lexicon, obtained from saifmohammad.com, serves as a pivotal resource for emotion analysis in natural language processing.
- The lexicon, encapsulated in the file "NRC-Emotion-Lexicon-Wordlevel-v0.92.txt," comprises an extensive collection of words associated with specific emotion categories.
- The lexicon includes a substantial vocabulary annotated with emotion categories, providing a comprehensive resource for sentiment and emotion analysis in natural language processing tasks.

DATA PREPROCESSING

Preprocessing of the data included several steps to ensure the quality and usability of the dataset:

- **Cleansing/Purging of Incomplete Data:** Tracks without essential information, such as missing audio features or artist details, were removed or corrected.
- **Aggregation and Summarization:** The data was consolidated at the track level. Each song's multiple attributes were compiled into a single record, facilitating easier analysis.
- **Handling Missing Values:** Checked the dataset for missing values and devised a strategy for handling them, opting for either removal of rows or imputation based on specific criteria.
- **Feature Engineering:** Introduced new features and modifications to extract meaningful information, such as extracting the release year from the date, calculating track duration, and deriving sentiment scores from lyrics.
- **Text Data Processing (Lyrics):** Implemented preprocessing for text data, specifically lyrics, by removing stop words, special characters, and converting text to lowercase.
- **Combining Data:** The main Spotify data and the lyrics data were merged based on track names and artist details. This integration allowed for a multi-dimensional analysis, considering both audio features and lyrical content.

METHODS OF ANALYSIS

1. **Top 50 Songs Analysis:**
    - **Questions to be Answered:**
    - What are the top 50 songs?
    - How many songs per artist are in the top 50?
    - What is the genre distribution in the top 50?
    - How is popularity distributed across tracks in the top 50?
    - What are the release years of the top 50 songs?
    - What is the sentiment analysis of the top 6 songs?
    - **Fields Used:**
        - Song details (name, artist, genre, popularity, release year)
        - Sentiment analysis of lyrics
    - **Collation of Results:**
        - **Genre Distribution Visualization:**
          Utilized Matplotlib to create a bar chart showcasing the distribution of genres within the top 50 songs. Each genre was represented proportionally to provide a visual overview of the diversity in musical styles.

- **Popularity Distribution Visualization:**
  Employed Matplotlib to generate a bar chart illustrating the distribution of popularity across the top 50 songs. This allowed for a quick assessment of the popularity range, emphasizing the most popular tracks.
- **Tabular Representation of Top 50 Songs:**
  Constructed a table using Pandas to display essential details of the top 50 songs, including song titles, respective artists, and their popularity. The tabular format allowed for easy reference and analysis of key song attributes.
- **Sentiment Analysis Presentation:**
  Obtained lyrics of the songs from genius API by providing the song name and the artist's name. Then performed Sentiment Analysis to the lyrics of the top songs, generating summary statistics to convey the prevalent sentiment. This offered insights into the emotional tone of the user's preferred tracks. Also, a pie chart was generated to show the percentage of positive and negative sentiments in the top 6 songs.

2. **Playlist Analysis:**
   - **Questions to be Answered:**
     - What is the genre of each playlist?
     - How does emotion vary across two playlists?
     - What are the audio features of tracks in the playlists?
     - How can a playlist be optimized for better danceability?
   - **Fields Used:**
     - Playlist details (name, genre)
     - Emotion analysis of playlist tracks
     - Audio features of playlist tracks (danceability)
   - **Collation of Results:**
     - **Genre Presentation for Each Playlist:**
       Employed Matplotlib to create individual bar charts showcasing the genre distribution within each playlist. Each bar represented a specific genre, offering a clear visual representation of the predominant musical styles in each playlist.
     - **Emotion Analysis Results for Specified Playlists:**
       Utilized NRC Emotion Lexicon for emotion analysis on lyrics, extracting dominant emotions for each track in a playlist for two playlists. Presented results through visualizations, such as bar charts, to convey the prevalent emotions in the selected playlists.

- **Statistical Analysis and Visualizations for Audio Features (Danceability):**
  Utilized statistical measures and visualizations, potentially with NumPy and Matplotlib, to analyze the danceability and valence audio feature across playlists. Suggestions for optimizations were provided based on the distribution of danceability scores, ensuring playlists align with user preferences for dance-friendly tracks.

3. **Optimization of the party Playlist to create a new playlist with higher Danceability:**
   - **Questions to be Answered:**
     - How can the 'Party' playlist be optimized to enhance the overall music experience and create a new playlist 'Party Vibes'?
     - What specific audio features are of primary concern for playlist optimization, considering potential trade-offs between features?
     - What is the impact of different approaches on danceability, valence, and an aggregated score of audio features?
   - **Fields Used:**
     - Audio features of playlist tracks
   - **Collation of Results:**
     - The optimization process for the 'Party' playlist involved meticulous consideration of various audio features to enhance the overall music experience.
     - Three distinct approaches were employed, each focusing on different aspects such as danceability, valence, and an aggregated score encompassing multiple features.
     - In Approach I, a careful sample of the 'Vibe' playlist was introduced based on danceability, resulting in an improved danceability mean but a decrease in valence.
     - Approach II refined the danceability further by filtering the 'Vibe' playlist based on specified criteria, yielding a playlist with increased danceability and valence.
     - Approach III introduced a comprehensive scoring system, considering danceability, energy, tempo, loudness, and valence, resulting in the creation of a new playlist that excelled in all specified criteria—higher danceability and valence, improved feature distributions, and the highest aggregate score.
     - The final playlist reflects a thoughtful blend of diverse tracks, providing a more engaging and uniform musical experience for the 'Party Vibe' playlist.

OVERALL DESCRIPTION OF THE PYTHON PROGRAM

The Python program is designed for comprehensive analysis of Spotify data, focusing on top songs and playlist insights. The program leverages various libraries such as Spotipy, Requests, NumPy, Pandas, Matplotlib, NLTK, and more.

Data Retrieval:

- Utilizes Spotipy and Requests to access Spotify Web API for fetching top songs and playlist details.

Data Processing:

- Cleans and preprocesses data, addressing missing values and ensuring consistency.

Exploratory Data Analysis (EDA):

- Performs EDA on the top 50 songs, examining artist distribution, genre composition, popularity, release years, and sentiment analysis of lyrics.

Playlist Analysis:

- Investigates playlist details, including genre, emotion analysis on specific playlists, and audio feature analysis for track optimization.

Sentiment and Emotion Analysis:

- Applies NLTK for sentiment and emotion analysis on lyrics, providing insights into the emotional content of songs.

Visualization:

- Utilizes Matplotlib for creating visual representations of data, facilitating easy interpretation of trends and patterns.

Optimization of the playlist:

- Proposes optimizations for playlist danceability based on analyzed audio features.

Methodology:

- Employs a systematic approach, including data cleaning, statistical analysis, and NLP techniques for sentiment and emotion assessment and then optimization of the playlist.

Collation of Results:

- Presents findings through visualizations (bar charts, pie charts), tabular summaries, and concise statements, ensuring accessibility and actionable insights.

Flexibility:

- Adaptable to different datasets, allowing users to explore and analyze Spotify data for various purposes.

User-Friendly Outputs:

- Generates clear and concise outputs, aiding users in making informed decisions about song preferences, playlist characteristics, and potential optimizations.

Overall, the Python program is a versatile tool for in-depth analysis of Spotify data, providing valuable insights for both casual users and music enthusiasts.

OUTPUT DOCUMENTATION

Top 50 Songs Analysis:

- **Top Artists:** List of top artists based on the number of songs in the top 50.
- **Genre Distribution:** Pie chart showcasing the distribution of genres in the top 50 songs.
- **Popularity Distribution:** Histogram illustrating the popularity distribution across tracks.
- **Release Year Distribution:** Bar chart displaying the release years of the top 50 songs.
- **Sentiment Analysis:** Emotional tone analysis of the lyrics in the top 6 songs.

Playlist Analysis:

- **Genre Overview:** Genre distribution summary for playlists.
- **Emotion Analysis:** Emotional content summary of specific playlists.
- **Audio Feature Analysis:** Visualization of audio features for track optimization.

Optimization Recommendations:

- **Danceability Metrics:** Comparison of danceability metrics for playlist optimization.
- **Recommendations:** Actionable recommendations for enhancing playlist danceability.

Visualizations:

- **Matplotlib Charts:** Clear visualizations for easy interpretation of data.
  Bar charts, pie charts, histograms, and line plots.

Tabular Summaries:

- **Top Artists Table:** Tabular representation of top artists and their song counts.
- **Genre Distribution Table:** Tabular summary of genre distribution in playlists.

Text Outputs:

- **Concise Statements:** Summarized textual statements describing key findings and insights.
- **Optimization Insights:** Clear statements providing insights into playlist optimization.

Flexibility and Interactivity:

- **User Prompts:** If applicable, prompts for user interactions or further analysis.
- **Dynamic Outputs:** Adaptability to different datasets and user queries.

CONCLUSIONS

- **Comprehensive Music Analysis:** This aspect of the project involves a thorough examination of various dimensions within music, encompassing both the inherent features of audio tracks and the lyrical content associated with them. By combining these two perspectives, a holistic understanding of the music is achieved.
- **User Music Preferences:** Through the analysis of playlists and individual tracks, the project delves into the user's specific preferences and inclinations when it comes to music. In this case, the user exhibits a pronounced liking for genres such as filmi and pop, as evidenced by the prevalence of these genres in their listening history.
- **Sentiment Analysis of Lyrics:** The application of VADER Sentiment Analysis on the lyrical content of songs provides valuable insights into the prevailing emotional tone of the user's preferred tracks. The analysis reveals a notable tendency towards songs with positive and uplifting sentiments.
- **Emotion Analysis with Lexicon:** Utilizing the NRC lexicon for emotion analysis adds a layer of depth to the understanding of the user's musical choices. This analysis highlights specific emotional tones in the preferred songs, indicating a preference for music that evokes joy and positivity while potentially avoiding more somber or melancholic themes.

GROUP CONTRIBUTION

The following questions were solved by each member:

1. **MALVIKA DIWAN**
    - What are the top 50 songs?
    - How many songs per artist are in the top 50?
    - What is the genre distribution in the top 50?
    - How is popularity distributed across tracks in the top 50?
2. **AYUSH SASEENDRAN**
    - What are the release years of the top 50 songs?
    - What is the sentiment analysis of the top 6 songs?
    - What is the distribution of positive and negative sentiments in the lyrics of the top 6 songs?
    - What is the genre of each playlist?
3. **KHUSHI SHETTY**
    - How does emotion vary across two playlists?
    - What are the audio features of tracks in the playlists?
    - How can the 'Party' playlist be optimized to enhance the overall music experience and create a new playlist ‘Party Vibes’?
    - What is the impact of different approaches on danceability, valence, and an aggregated score of audio features?

The presentation and report were collaboratively crafted by all team members.