Spotify Music Data Analysis – Project Documentation

1. Project Overview

The Spotify Music Data Analysis project focuses on analyzing a large Spotify dataset to understand song characteristics and their relationship with popularity. Spotify, being one of the world's leading music streaming platforms, provides rich audio feature data such as danceability, energy, tempo, loudness, and valence. This project applies data analysis and basic machine learning techniques to extract meaningful insights from these features and to predict song popularity.

2. Objectives of the Project

The main objectives of this project are: - To analyze various audio features of songs such as danceability, energy, tempo, loudness, and valence. - To understand how these audio features influence song popularity on Spotify. - To perform data cleaning and preprocessing to ensure data quality. - To visualize trends, distributions, and relationships in the data. - To build and evaluate predictive models for forecasting song popularity.

3. Dataset Description

The dataset used in this project contains Spotify track-level data in CSV format. Each row represents a single song, and each column represents a specific attribute of that song.

Key attributes in the dataset include:

- **Basic song information:** track name, artist(s), release year, release date.
- **Audio features:** danceability, energy, tempo, loudness, acousticness, instrumentalness, speechiness, liveness, and valence.
- **Popularity:** a numerical score (0–100) indicating how popular a song is on Spotify.

The dataset contains over **170,000 songs** with **48 features**, making it suitable for large-scale exploratory and predictive analysis.

4. Tools and Technologies Used

- **Programming Language:** Python
- **Libraries:**
    - Pandas – data manipulation and analysis
    - NumPy – numerical computations
    - Matplotlib & Seaborn – data visualization
    - Scikit-learn – machine learning models and evaluation

- **Environment:** Jupyter Notebook / Google Colab

5. Data Loading and Preprocessing

5.1 Data Loading

The dataset was loaded into the Python environment using Pandas. Initial inspection was performed to understand the structure, size, and data types of the columns.

5.2 Data Cleaning

The following preprocessing steps were applied: - Duplicate rows were removed to avoid biased analysis. - Missing values in numerical columns were handled using **mean imputation**. - Data types were verified to ensure correct numerical and categorical handling.

These steps ensured that the dataset was clean and ready for analysis without losing significant information.

6. Exploratory Data Analysis (EDA)

EDA was conducted to understand the distribution and behavior of different features.

6.1 Descriptive Statistics

- Summary statistics such as mean, median, minimum, maximum, and standard deviation were calculated for numerical features.
- This helped identify the overall range and variability of audio features.

6.2 Feature Distributions

- Histograms were plotted for features like danceability, energy, tempo, and popularity.
- The popularity distribution showed that most songs have low to moderate popularity, with fewer highly popular songs.

6.3 Trends Over Time

- Average song popularity was analyzed across years.
- The trend shows a gradual increase in average popularity over time, indicating changes in music consumption and listener behavior.

6.4 Correlation Analysis

- Correlation matrices and scatter plots were used to analyze relationships between features.
- A positive correlation was observed between **energy and danceability**, indicating that energetic songs are often more danceable.

7. Data Visualization

Multiple visualizations were created to represent insights clearly: - **Bar charts:** Average popularity by year. - **Line graphs:** Trends of popularity and energy over time. - **Scatter plots:** Danceability vs popularity, energy vs danceability. - **Heatmaps:** Correlation between key audio features. - **Pair plots:** Multivariate relationships between selected features.

These visualizations helped in identifying trends, patterns, and relationships in an intuitive manner.

8. Modeling and Prediction

8.1 Feature Selection

The following features were selected as predictors: - Danceability - Energy - Tempo - Valence - Loudness

The target variable was **song popularity**.

8.2 Model Building

Two machine learning models were implemented: - **Linear Regression** – as a baseline model. - **Decision Tree Regressor** – to capture non-linear relationships.

The dataset was split into training (80%) and testing (20%) sets.

8.3 Model Evaluation

Models were evaluated using: - **Mean Squared Error (MSE)** - **R^2 Score**

Results showed: - Linear Regression achieved moderate performance. - Decision Tree initially performed worse but improved significantly after hyperparameter tuning.

8.4 Model Tuning

The Decision Tree model was fine-tuned using parameters such as maximum depth and minimum samples split, resulting in improved prediction accuracy.

9. Key Findings

- Popular songs tend to have **higher energy and danceability**.
- Song popularity has increased over time, reflecting evolving listener preferences.
- Energy and danceability show a positive correlation.
- Machine learning models can reasonably predict song popularity using audio features.
- The tuned Decision Tree model performed better than Linear Regression.

10. Implications of the Results

- **Artists and producers** can focus on energetic and danceable tracks to improve popularity.
- **Music platforms** can enhance recommendation systems using feature-based insights.
- Predictive modeling can help forecast a song's success before release.

11. Limitations

- Genre-level analysis was limited due to missing genre values.
- Popularity is influenced by external factors (marketing, artist fame) not included in the dataset.
- Models used were basic and can be improved further.

12. Future Scope

- Perform genre-based popularity analysis.
- Include artist-level popularity and listener behavior data.
- Apply advanced models such as Random Forest, Gradient Boosting, or Neural Networks.
- Analyze lyrics and sentiment for deeper insights.
- Implement real-time streaming data analysis.

13. Conclusion

This project successfully demonstrated how data analysis and machine learning can be applied to music data to extract meaningful insights. Through systematic preprocessing, exploratory analysis, visualization, and modeling, the project highlights key factors influencing song popularity and provides a strong foundation for more advanced music analytics in the future.