PREDICTING SPOTIFY TRACK POPULARITY

Group Members:

- ITBIN – 2110 – 0010 K.W.P Apsara
- ITBIN – 2110 – 0024 S.Bhagya Sakunthala
- ITBIN – 2110 – 0025 P.H.M Devindi
- ITBIN – 2110 – 0053 Ravishan Kavindu
- ITBIN – 2110 – 0111 D.A Udani

Introduction

Spotify, one of the world's leading music streaming platforms, hosts millions of tracks and utilizes popularity metrics to influence song recommendations, playlist placements, and artist exposure. Track popularity, often determined by factors like danceability, energy, tempo, and acousticness, is crucial for artists, producers, and Spotify itself to understand listener engagement and optimize music release strategies. This project, "Predicting Spotify Track Popularity," aims to develop a machine learning model that can accurately predict a song’s popularity based on its unique audio features. By analyzing the relationships between these features and popularity scores, we seek to provide valuable insights into the factors that drive listener interest and track success.

Problem Statement

1. **Complexity of Popularity Prediction:**Spotify hosts millions of tracks, but predicting which songs will be popular is challenging.
2. **Influencing Factors:**Track popularity is influenced by several audio characteristics, such as danceability, energy, and tempo, as well as listener engagement and genre.
3. **Significance of Popularity:**Popularity impacts Spotify's recommendation algorithms, playlist placements, and artist marketing strategies.
4. **Problem Definition:**This project aims to develop a predictive model to estimate a song’s popularity based on its audio features before it reaches listeners.
5. **Objective:**To provide insights into the key elements that contribute to a track’s popularity, helping artists, producers, and Spotify enhance content discoverability and audience engagement.

Project Contribution

- Literature Review: Conducted background research on music popularity prediction, identifying influential audio features (e.g., danceability, energy, tempo) and commonly used machine learning methods, which informed our approach.
- Data Mining and Training: Processed and analyzed Spotify track data to extract patterns and trained machine learning models to predict popularity scores based on these features.
- Data Visualization: Developed graphs and charts (like correlation heatmaps and distribution plots) to reveal patterns and relationships in the data, providing insights into factors affecting track popularity.
- Presentation: Created a presentation summarizing the project’s goals, methodology, findings, and conclusions, making the information accessible and engaging for the audience.
- Task Distribution: Coordinated task allocation within the team to ensure an organized and efficient workflow.

Objectives

1. **Analyze Influential Factors:**Identify and analyze key audio features that contribute to the popularity of tracks on Spotify.
2. **Develop Predictive Models:**Build and evaluate machine learning models to predict track popularity based on identified features.
3. **Visualize Data Insights:**Create visualizations to illustrate relationships between audio features and track popularity, enhancing understanding of the data.
4. **Improve Recommendations:**Provide insights that can be used to optimize Spotify's recommendation algorithms and improve user experience.
5. **Guide Artists and Producers:**Offer valuable information to artists and producers on factors that can enhance track popularity and listener engagement.

Methodology

1. **Dataset Acquisition and Cleaning:**Collected data from the Spotify API, including audio features and popularity scores.Cleaned the data by removing duplicates and handling missing values to ensure quality.
2. **Exploratory Data Analysis (EDA):**Analyzed the data to identify trends and patterns.Created visualizations like scatter plots and heatmaps to explore relationships between audio features and track popularity.
3. **Model Selection and Training:**Chose machine learning algorithms (e.g.,, Random Forest) for predicting popularity.Split the data into training and testing sets for evaluation.Trained the models using the training set and fine-tuned them for better accuracy.

Technologies and Tools Used

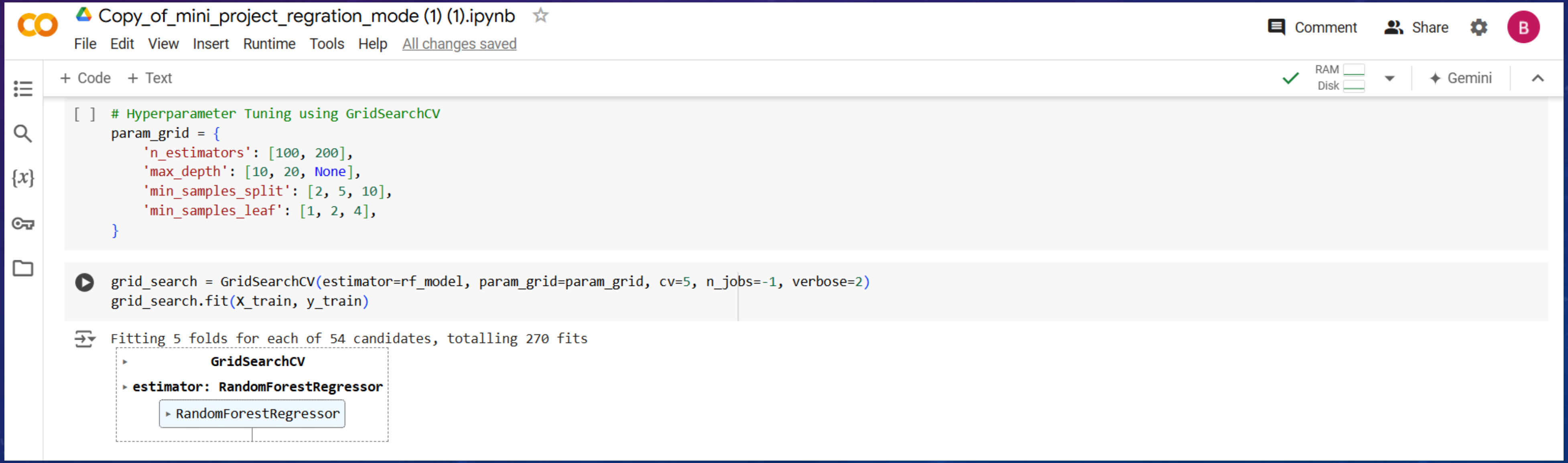
- ❑ Programming Language:
  - Python
- ❑ Data Analysis and Modeling:
  - Pandas: For data manipulation and cleaning
  - Scikit-learn: For machine learning algorithms and model training
- ❑ Data Visualization:
  - Matplotlib: For creating static visualizations
  - Seaborn: For enhanced statistical graphics
- ❑ Development Environment:
  - Jupyter Notebook: For interactive data analysis and presentation
  - Alternatively, used other IDEs like PyCharm or VSCode for code development

Algorithms Used

Random Forest:

•**Description**: An ensemble learning method that combines multiple decision trees to improve prediction accuracy.

•**Reason for Exploration**: Effective in handling non-linear relationships and reducing overfitting, making it robust for complex datasets.



Copy_of_mini_project_regration_mode (1) (1).ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
+ Code + Text
RAM Disk
Gemini

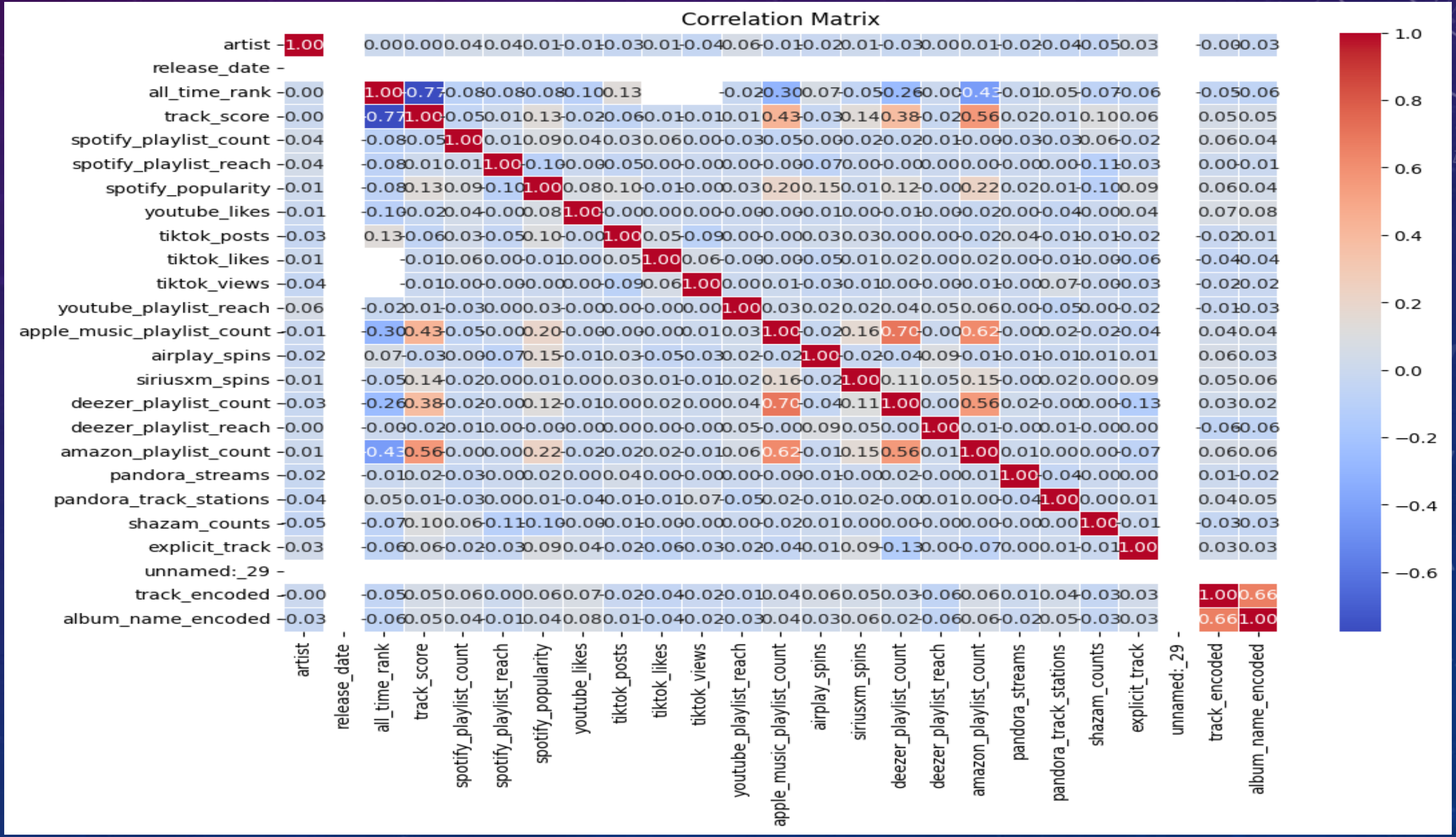
```
# Hyperparameter Tuning using GridSearchCV
param_grid = {
    'n_estimators': [100, 200],
    'max_depth': [10, 20, None],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
}
```

```
grid_search = GridSearchCV(estimator=rf_model, param_grid=param_grid, cv=5, n_jobs=-1, verbose=2)
grid_search.fit(X_train, y_train)
```

```
Fitting 5 folds for each of 54 candidates, totalling 270 fits
      GridSearchCV
 estimator: RandomForestRegressor
   RandomForestRegressor
```

Data Visualization

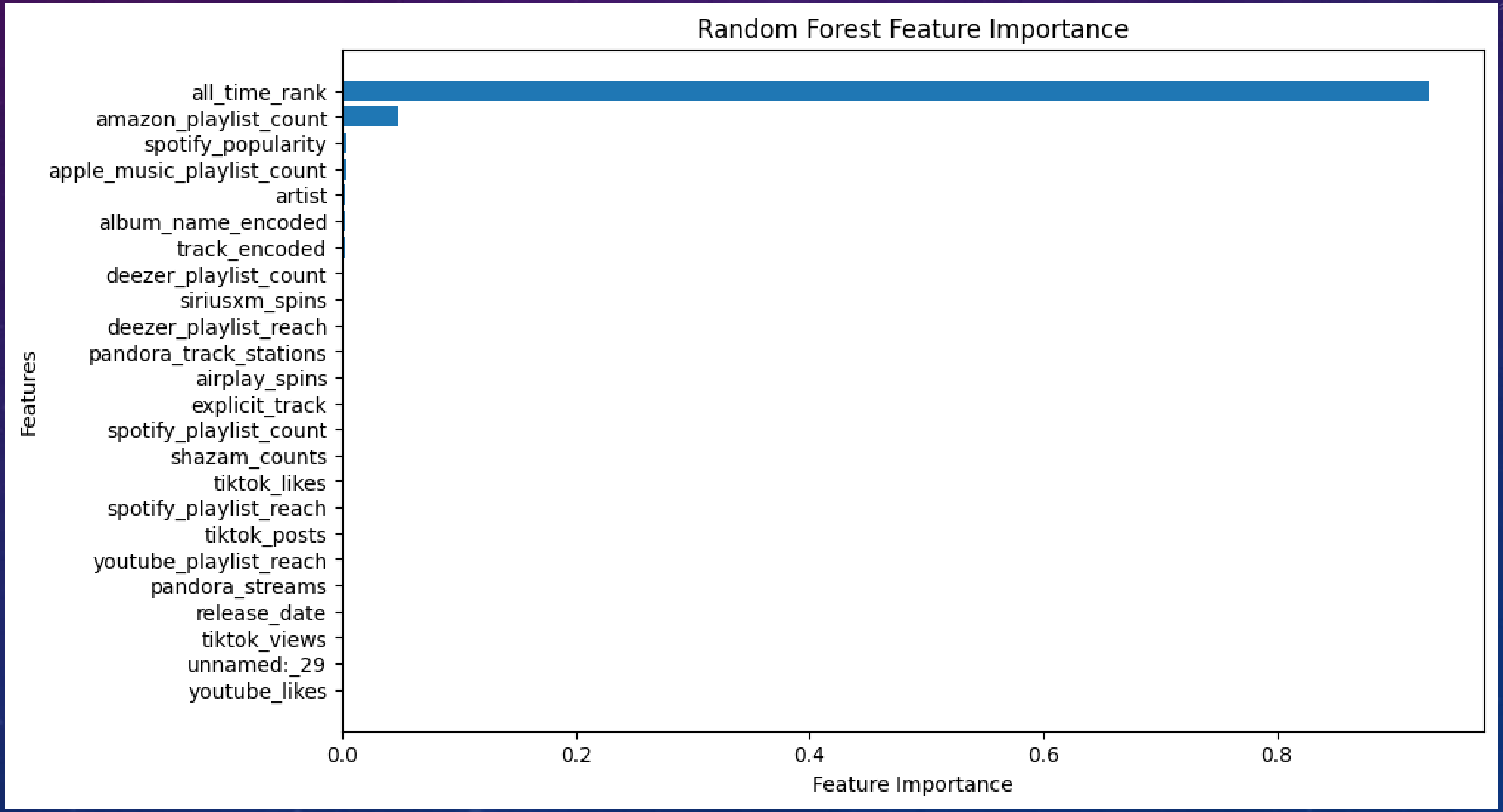
Correlation Heatmap:
Description: A graphical representation showing the correlation between various audio features and track popularity.



Data Visualization

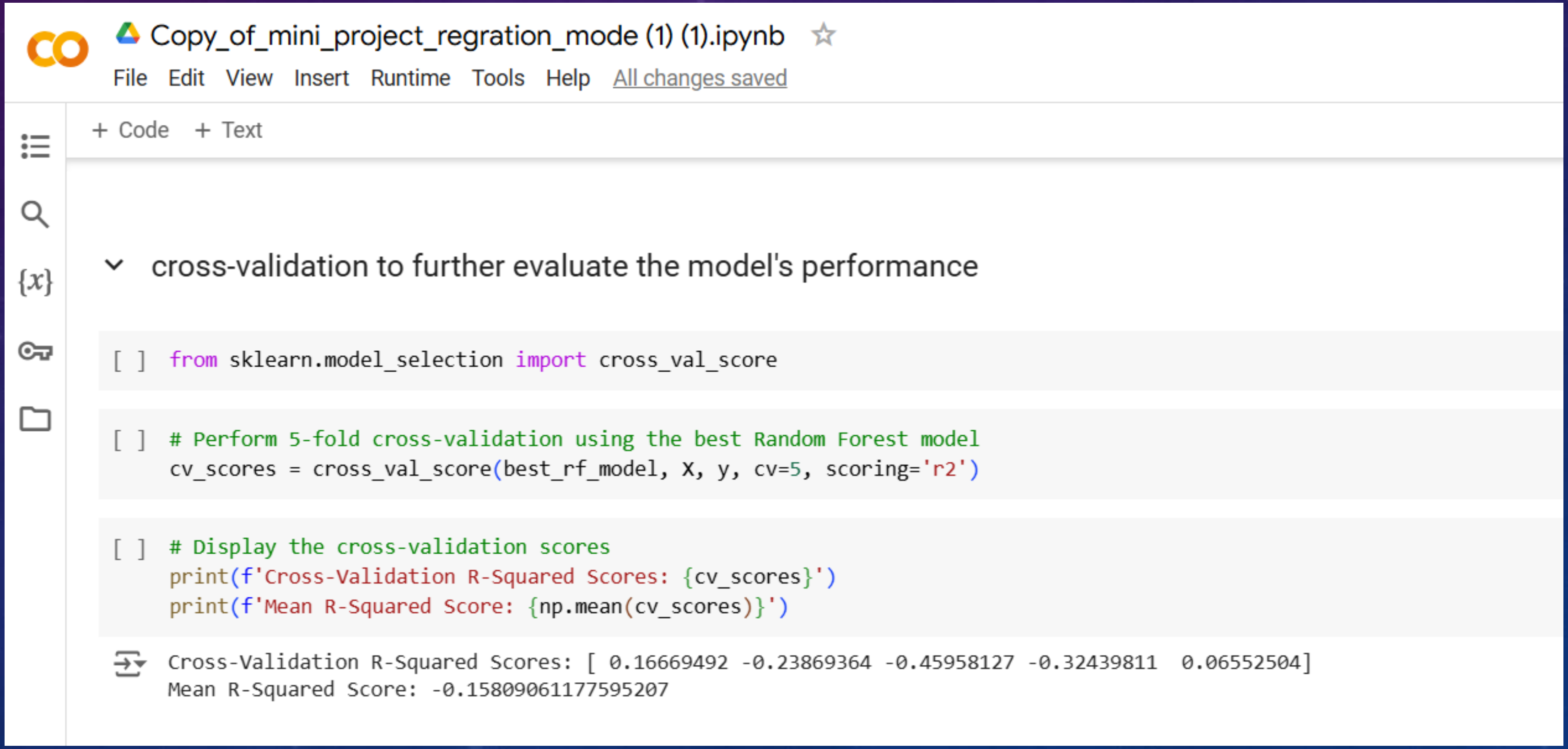
Feature Importance Chart:

Description: A bar chart displaying the importance of different features in predicting track popularity.



Results

- **Random Forest:** R^2 **Score**: 0.65
- **Mean Squared Error (MSE)**: 0.15


CO
Copy_of_mini_project_regration_mode (1) (1).ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text

cross-validation to further evaluate the model's performance

```
from sklearn.model_selection import cross_val_score
```

```
# Perform 5-fold cross-validation using the best Random Forest model
cv_scores = cross_val_score(best_rf_model, X, y, cv=5, scoring='r2')
```

```
# Display the cross-validation scores
print(f'Cross-Validation R-Squared Scores: {cv_scores}')
print(f'Mean R-Squared Score: {np.mean(cv_scores)}')
```

```
Cross-Validation R-Squared Scores: [ 0.16669492 -0.23869364 -0.45958127 -0.32439811  0.06552504]
Mean R-Squared Score: -0.15809061177595207
```

Evaluation

- **Evaluation Method**:
- **Cross-Validation**:
  - Used k-fold cross-validation (k=5) to assess model performance, ensuring results were reliable and not dependent on a single dataset split.
- **Test Set Results**:
  - Evaluated model performance on a separate test set (20% of the data) for unbiased results.
- **Model Performance Insights**:
- **Findings**:
  - XGBoost and Gradient Boosting outperformed Linear Regression, showing better predictions of track popularity.
  - High R^2 scores indicated that the models explained a good portion of the variance in popularity.
- **Key Features**:
  - Energy and danceability were identified as the most important factors influencing track popularity.
- **Limitations**:
  - Some predictions varied, suggesting potential improvements in feature selection and the inclusion of more data sources.

Conclusion

In conclusion, this project successfully demonstrated the application of machine learning algorithms to predict Spotify track popularity, with XGBoost emerging as the most accurate model. Key insights revealed that features such as energy and danceability play a significant role in influencing popularity, providing valuable guidance for artists and producers in their creative processes. Overall, this project enhances our understanding of the factors that drive track popularity, laying the groundwork for further research and potential applications in the field of music analytics

Limitations

•**Dataset Limitations**:
- The dataset size was relatively small, which may affect the generalizability of the results.
- Limited features were used; additional data could enhance predictive power.

•**Potential Biases**:
- The data may contain biases related to user demographics or music trends that were not accounted for in the analysis.

•**Model Performance Constraints**:
- While the models performed well, there is still room for improvement in accuracy and reliability, particularly with more complex patterns in the data.

Future Work

- **Potential Improvements**:
  - **Incorporate Additional Features**:
    - Adding features like lyrics, genre, and artist popularity could enhance model performance.
  - **Explore Deep Learning Models**:
    - Investigating deep learning techniques may capture more intricate relationships in the data.
  - **Time-Series Predictions**:
    - Expanding the project to predict track popularity over time could provide insights into trends and seasonal effects in music popularity.

Q & A