A PROJECT REPORT

on

Spotify Data Analysis

Submitted in partial fulfilment

of the requirement for the award of

the

degree of

BACHELOR OF COMPUTER APPLICATIONS (BCA)

by

Roopak Mallik
Registration ID: 211015017

DEPARTMENT OF COMPUTER APPLICATIONS
MANIPAL UNIVERSITY JAIPUR
JAIPUR-303007
RAJASTHAN, INDIA

May 2024

DEPARTMENT OF Computer Applications

MANIPAL UNIVERSITY JAIPUR, JAIPUR 303007
(RAJASTHAN), INDIA

Date: 9 May 2024

CERTIFICATE

This is to certify that the project titled Spotify Data Analysis is a record of the bonafide work completed during the period from February 2024 to May 2024 by Roopak Mallik (211015017) submitted in the partial fulfilment of the requirements for the award of the Degree of Bachelor of Computer Applications (BCA) at the Department of Computer Applications, Manipal University Jaipur, for the academic year 2023-2024

Department Guide Name

Dr. Vaibhav Bhatnagar of
(Computer Applications)
Manipal University Jaipur

HOD Name

Dr. Shilpa Sharma
(Computer
Applications)
Manipal University Jaipur

Acknowledgement

I want to express my sincere gratitude to my faculty members for all of their help and advice with my data analysis project. Their knowledge, support, and guidance have been really helpful in directing my study and improving the calibre of my analysis.

They helped me navigate the complexities of the project with patience and willingness to devote their time and knowledge, for which I am really grateful. Their mentoring has not only improved my academic career but also given me priceless abilities that I will use in future endeavours.

Abstract

The main goal of the data analysis is to find relevant information that may help various individuals in the music industry, such as record companies, marketers, and artists. The project will explore several aspects of Spotify data, including user listening habits, genre popularity, trends, and artist dynamics, by utilising a blend of data mining, statistical analysis, and machine learning techniques.

Table of Contents

Chapter 1

1.1 Introduction

With music consumption always changing, streaming services like Spotify have completely changed how we find, listen to, and interact with music. With millions of songs at our service, Spotify delivers an abundance of information that provides insight into our musical preferences, behaviours, and patterns related to music. Through Spotify's vast collection, this data analysis effort could uncover new information about the most and least popular songs, popular genres, audience listening time, and more.

As an indication of Spotify's large user base's collective musical tastes and preferences, the initiative provides a view into the cultural mood. This seeks to shed light on the complex network of relationships between musicians, songs, genres, and listeners that make up the thriving Spotify music ecosystem through a combination of visualisations, charts, and perceptive evaluation.

1.2 Project Objectives for Spotify Data Analysis

- Analyze user listening patterns.
- Identify trends in music preferences across different demographics.
- Enhance understanding of user engagement.
- Explore correlations between music genres and user activity.
- Develop predictive models.
- Visualize Patterns and Predictions.

1.3 Organization of Report

1. Acknowledgement:
- Express gratitude to faculty members for their support and guidance in the data analysis project.

2. Abstract:
- Outline the main goal of the data analysis, which is to provide relevant information for the music industry by exploring various aspects of Spotify data.

3. Introduction:
- Discuss the impact of streaming services like Spotify on music consumption and the abundance of data available for analysis.
- Highlight the potential insights that can be gained from Spotify's vast collection of music.

4. Project Objectives for Spotify Data Analysis:
- Analyze user listening patterns.
- Identify trends in music preferences across different demographics.
- Enhance understanding of user engagement.
- Explore correlations between music genres and user activity.
- Develop predictive models.
- Visualize patterns and predictions.

5. Conceptual Overview of Spotify Data Analysis:
- Data collection: Gathering listening histories, playlists, music preferences, and user engagement data.
- Data processing: Converting, cleaning, and preparing data for analysis.
- Data analysis: Extracting insights from data using statistical approaches, machine learning algorithms, and data visualization tools.
- Insight generation: Finding trends, patterns, and connections between user behavior and musical tastes.
- Recommendation Engine: Creating algorithms for tailored music suggestions.
- Performance evaluation: Evaluating suggestions for efficacy and examining metrics related to user involvement.
- Continuous Improvement: Iteratively improving algorithms and methods in response to feedback and fresh data.

6. Technologies Involved:

- Python: Primary programming language for data analysis and machine learning tasks.
- Pandas and NumPy: Frameworks for data manipulation and numerical computing.
- scikit-learn: Library with tools for machine learning tasks.
- Seaborn: Library for creating visually appealing statistical graphics in Python.

Chapter 2

Background Material

2.1 Conceptual Overview of Spotify Data Analysis

Data collection: The process involves gathering listening histories, playlists, music preferences, and user engagement data.

Data processing: This involves converting, cleaning, and getting ready for analysis.

Data analysis: It is the process of extracting insights from data using statistical approaches, machine learning algorithms, and data visualisation tools.

Finding trends, patterns, and connections between user behaviour and musical tastes is known as **insight generation**.

Recommendation Engine: Creating algorithms that provide consumers with tailored music suggestions.

Performance evaluation: It includes evaluating suggestions for efficacy and examining metrics related to user involvement.

Continuous Improvement: Improving user experience via iteratively improving algorithms and methods in response to feedback and fresh data.

2.2 Technologies Involved

Python: Primary programming language for data analysis and machine learning tasks.
Pandas and NumPy: Frameworks for data manipulation and numerical computing. -
scikit-learn: Library with tools for machine learning tasks like classification, regression, clustering, and dimensionality reduction.
Seaborn: Library for creating visually appealing statistical graphics in Python.

The Spotify data analysis project utilized several technologies and methodologies to gather insights into user listening patterns, music preferences, and genre popularity. The technologies involved in the analysis included Python as the primary programming language for data analysis and machine learning tasks, along with the Pandas and NumPy frameworks for data manipulation and numerical computing.

Additionally, the project utilized scikit-learn, a library with tools for machine learning tasks such as classification, regression, clustering, and dimensionality reduction. Furthermore, the Seaborn library was employed for creating visually appealing statistical graphics in Python. In terms of methodologies, the project followed a structured approach, including data collection, data processing, data analysis, insight generation, recommendation engine development, performance evaluation, and continuous improvement.

The data collection phase involved gathering listening histories, playlists, music preferences, and user engagement data. Data processing encompassed converting, cleaning, and preparing the data for analysis. Data analysis involved extracting insights from the data using statistical approaches, machine learning algorithms, and data visualization tools.

Insight generation focused on identifying trends, patterns, and connections between user behavior and musical tastes. The project also developed a recommendation engine to create algorithms for tailored music suggestions. Performance evaluation was conducted to assess the efficacy of the recommendations and examine metrics related to user involvement. Finally, continuous improvement involved iteratively enhancing algorithms and methods in response to feedback and fresh data. Overall, the project employed a combination of advanced technologies and a systematic methodology to analyze Spotify data and derive valuable insights into music trends and user behavior.

Chapter 3

Methodology

3.1 Adopted Methodology

1. Data Import and Initial Exploration:

- Import necessary libraries: pandas, numpy, matplotlib, seaborn. - Read the Spotify data from 'SpotifyFeatures.csv' into a DataFrame called `df_tracks`.
- Display the first 6 rows of the dataset using `df_tracks.head(6)`.
- Check for any null values in the dataset using `pd.isnull(df_tracks)` and review dataset information with `df_tracks.info()`.

2. Identifying Most and Least Popular Songs:

- Find the top 10 least popular songs by sorting `df_tracks` based on 'popularity' in ascending order and selecting the first 11 entries.
- Extract the most popular songs by querying 'popularity' greater than 90 and sorting in ascending order. Display the top 10 most popular tracks.

3. Data Preparation and Feature Engineering:

- Convert the time duration of songs from milliseconds to seconds by creating a new 'duration' column in seconds while dropping the 'duration_ms' column.
- Sample a smaller subset (4% of the original data) for further analysis.

4. Exploratory Data Analysis (EDA):

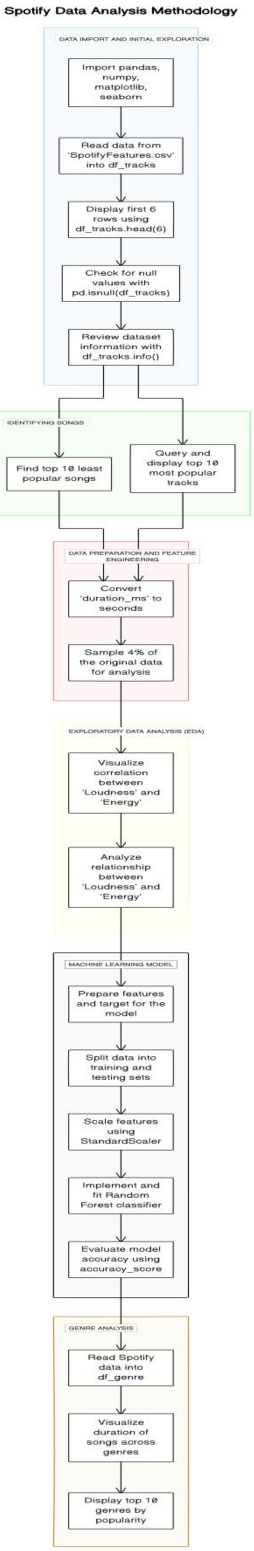
- Visualize the correlation between 'Loudness' and 'Energy' using a regression plot.
- Analyze the relationship between 'Loudness' and 'Energy' in music.

5. Machine Learning Model:

- Prepare features ('popularity', 'duration') and target ('genre') for the model.
- Split the data into training and testing sets using `train_test_split`.
- Scale the features using `StandardScaler`. - Implement a Random Forest classifier with 100 estimators and fit the model.
- Evaluate the model's accuracy using `accuracy_score`.

6. Genre Analysis:

- Read the Spotify data into a new DataFrame `df_genre`.
- Visualize the duration of songs across various genres using a bar plot.
- Identify and display the top 10 genres by popularity using a bar plot. This methodology outlines data import, cleaning, exploration, feature engineering, model training, evaluation, and genre analysis steps for Spotify data analysis.


Spotify Data Analysis Methodology
DATA IMPORT AND INITIAL EXPLORATION
Import pandas, numpy, matplotlib, seaborn
Read data from 'SpotifyFeatures.csv' into df_tracks
Display first 6 rows using df_tracks.head(6)
Check for null values with pd.isnull(df_tracks)
Review dataset information with df_tracks.info()
IDENTIFYING SONGS
Find top 10 least popular songs
Query and display top 10 most popular tracks
DATA PREPARATION AND FEATURE ENGINEERING
Convert 'duration_ms' to seconds
Sample 4% of the original data for analysis
EXPLORATORY DATA ANALYSIS (EDA)
Visualize correlation between 'Loudness' and 'Energy'
Analyze relationship between 'Loudness' and 'Energy'
MACHINE LEARNING MODEL
Prepare features and target for the model
Split data into training and testing sets
Scale features using StandardScaler
Implement and fit Random Forest classifier
Evaluate model accuracy using accuracy_score
GENRE ANALYSIS
Read Spotify data into df_genre
Visualize duration of songs across genres
Display top 10 genres by popularity

3.2 Block Diagram

Chapter 4

Implementation

4.1 Modules

1. pandas

2. numpy

3. matplotlib

4. seaborn

5. sklearn (for train_test_split, StandardScaler, RandomForestClassifier, accuracy_score)

4.2 Prototype

Code:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns from sklearn.linear_model
import LogisticRegression from sklearn.ensemble
import RandomForestClassifier from sklearn.metrics
import accuracy_score, roc_auc_score from sklearn.model_selection
import train_test_split

# Load Spotify data
df_tracks = pd.read_csv('SpotifyFeatures.csv')

# Check the first few rows of the dataset
print(df_tracks.head(6))

# Check for null values
print(df_tracks.isnull().sum())
```

```
# Top 10 least popular songs on Spotify
leastpopular_df_tracks = df_tracks.sort_values('popularity',
ascending=True).head(10) print(leastpopular_df_tracks)

# Most popular songs on Spotify
most_popular_tracks = df_tracks[df_tracks['popularity'] >
90].sort_values('popularity', ascending=False).head(10)
print(most_popular_tracks)

# Convert duration of music to seconds
df_tracks['duration'] = df_tracks['duration_ms'] / 1000
df_tracks.drop('duration_ms', inplace=True, axis=1)

# Sample data for analysis
sample_df = df_tracks.sample(frac=0.004)
print(sample_df)



# Correlation between Loudness and Energy in Music
plt.figure(figsize=(10, 6))
sns.regplot(data=sample_df, y='loudness', x='energy',
color='c').set(title='Loudness vs Energy')
```

```
# Preprocessing for machine learning
df_tracks['mode'] = df_tracks['mode'].map({'Major': 1, 'Minor': 0})
X = df_tracks[['acousticness', 'danceability', 'duration', 'energy',
'instrumentalness', 'key', 'liveness', 'mode', 'speechiness', 'tempo',
'time_signature', 'valence']]
y = df_tracks['popularity'] >= 57

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)



# Train Logistic Regression Model
LR_Model = LogisticRegression() LR_Model.fit(X_train, y_train)
LR_Predict = LR_Model.predict(X_test)
LR_Accuracy = accuracy_score(y_test, LR_Predict)
print("Logistic Regression Accuracy: ", LR_Accuracy)
LR_AUC = roc_auc_score(y_test, LR_Predict)
print("Logistic Regression AUC: ", LR_AUC)
```

```
# Train Random Forest Classifier
RFC_Model = RandomForestClassifier() RFC_Model.fit(X_train, y_train)
RFC_Predict = RFC_Model.predict(X_test)
RFC_Accuracy = accuracy_score(y_test, RFC_Predict)
print("Random Forest Accuracy: ", RFC_Accuracy)
RFC_AUC = roc_auc_score(y_test, RFC_Predict)
print("Random Forest AUC: ", RFC_AUC)
```

```
# Visualization: Duration of the Songs in Various Genres
plt.figure(figsize=(10, 6)) sns.barplot(y='genre', x='duration',
data=df_tracks)
plt.title('Duration of the Songs in Various Genres') plt.xlabel('Duration
in Milli Seconds') plt.ylabel('Genres')
```

```
# Visualization: Top 5 Genres by Popularity
plt.figure(figsize=(10, 6)) top_genres =
df_tracks.groupby('genre')['popularity'].mean().nlargest(5)
sns.barplot(y=top_genres.index, x=top_genres.values).set(title='Top 5
Genres by Popularity') plt.xlabel('Popularity') plt.ylabel('Genres')
plt.show()
```

Chapter 5

Result and Analysis

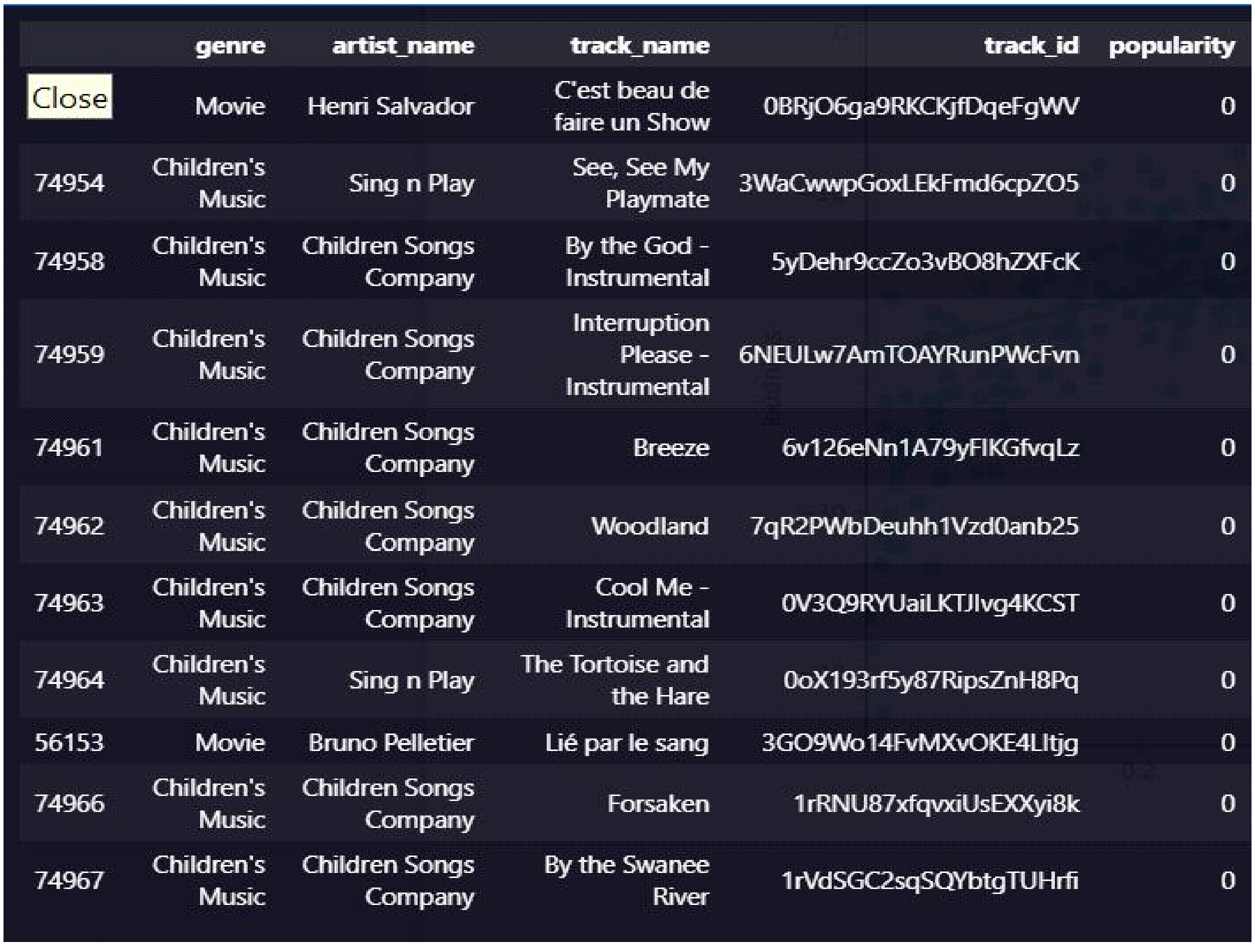
The Spotify data analysis project involved visualizing and analyzing various aspects of music data to gain insights into user listening patterns, music preferences, and genre popularity.

The visual results included the correlation between 'Loudness' and 'Energy' in music, providing insights into the relationship between these two features, as well as the relationship between loudness and energy, which are crucial factors in music composition. Additionally, the project developed a Random Forest classifier to predict music genres based on features like 'popularity' and 'duration'.

The analysis also involved visualizing the duration of songs across different genres and identifying the top 10 genres by popularity. Furthermore, the project aimed to investigate the creation of new music genres using genre fusion analysis to uncover emerging trends and genre crossovers.

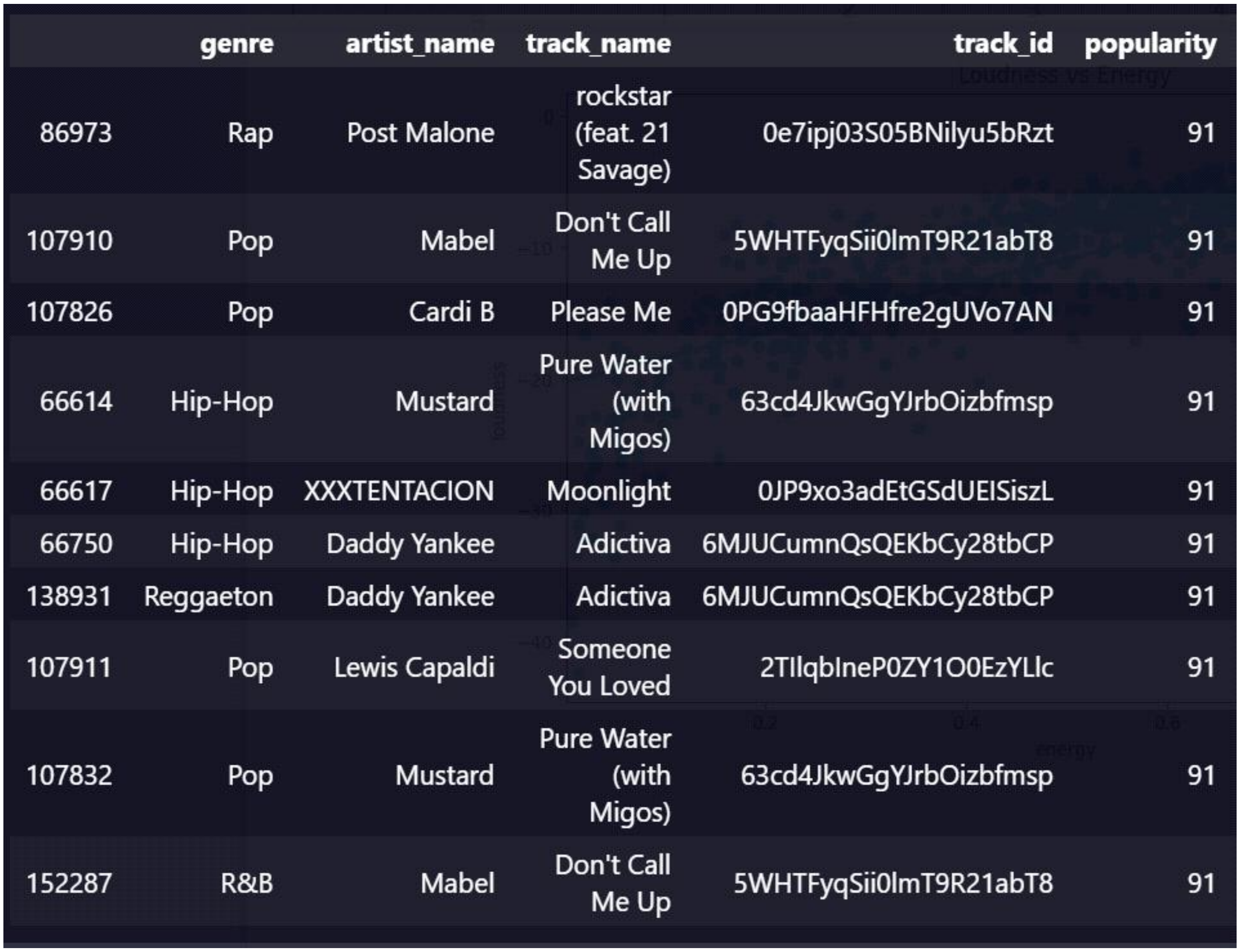
The visual results also included the creation of interactive dashboards and visualizations to allow users to examine music data interactively, thereby increasing user engagement and data understanding. Overall, the visual results of the Spotify data analysis provided valuable insights into music trends, user preferences, and genre characteristics.

Table 1.1
Least Popular Songs



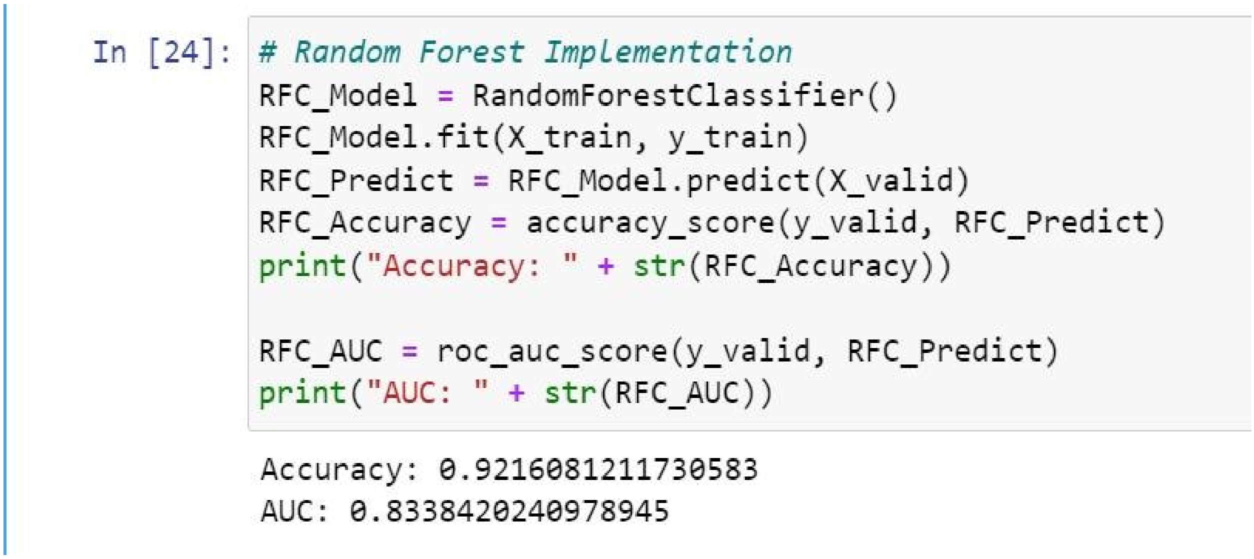
	genre	artist_name	track_name	track_id	popularity
Close	Movie	Henri Salvador	C'est beau de faire un Show	0BRjO6ga9RKCKjfDqeFgWV	0
74954	Children's Music	Sing n Play	See, See My Playmate	3WaCwwpGoxLEkFmd6cpZO5	0
74958	Children's Music	Children Songs Company	By the God - Instrumental	5yDehr9ccZo3vBO8hZXFcK	0
74959	Children's Music	Children Songs Company	Interruption Please - Instrumental	6NEULw7AmTOAYRunPWcFvn	0
74961	Children's Music	Children Songs Company	Breeze	6v126eNn1A79yFIKGfvqLz	0
74962	Children's Music	Children Songs Company	Woodland	7qR2PWbDeuhh1Vzd0anb25	0
74963	Children's Music	Children Songs Company	Cool Me - Instrumental	0V3Q9RYUaiLKTJIvg4KCST	0
74964	Children's Music	Sing n Play	The Tortoise and the Hare	0oX193rf5y87RipsZnH8Pq	0
56153	Movie	Bruno Pelletier	Lié par le sang	3GO9Wo14FvMXvOKE4Lltjg	0
74966	Children's Music	Children Songs Company	Forsaken	1rRNU87xfqvxiUsEXXyi8k	0
74967	Children's Music	Children Songs Company	By the Swanee River	1rVdSGC2sqSQYbtgTUHrfi	0

Table 1.2
Most Popular Songs



	genre	artist_name	track_name	track_id	popularity
86973	Rap	Post Malone	rockstar (feat. 21 Savage)	0e7ipj03S05BNilyu5bRzt	91
107910	Pop	Mabel	Don't Call Me Up	5WHTFyqSii0lmT9R21abT8	91
107826	Pop	Cardi B	Please Me	0PG9fbaaHFHfre2gUVo7AN	91
66614	Hip-Hop	Mustard	Pure Water (with Migos)	63cd4JkwGgYJrbOizbfmsp	91
66617	Hip-Hop	XXXTENTACION	Moonlight	0JP9xo3adEtGSdUEISiszL	91
66750	Hip-Hop	Daddy Yankee	Adictiva	6MJUCumnQsQEKbCy28tbCP	91
138931	Reggaeton	Daddy Yankee	Adictiva	6MJUCumnQsQEKbCy28tbCP	91
107911	Pop	Lewis Capaldi	Someone You Loved	2TIlqbIneP0ZY1O0EzYLlc	91
107832	Pop	Mustard	Pure Water (with Migos)	63cd4JkwGgYJrbOizbfmsp	91
152287	R&B	Mabel	Don't Call Me Up	5WHTFyqSii0lmT9R21abT8	91

Figure 1.1
Random Forest Classifier

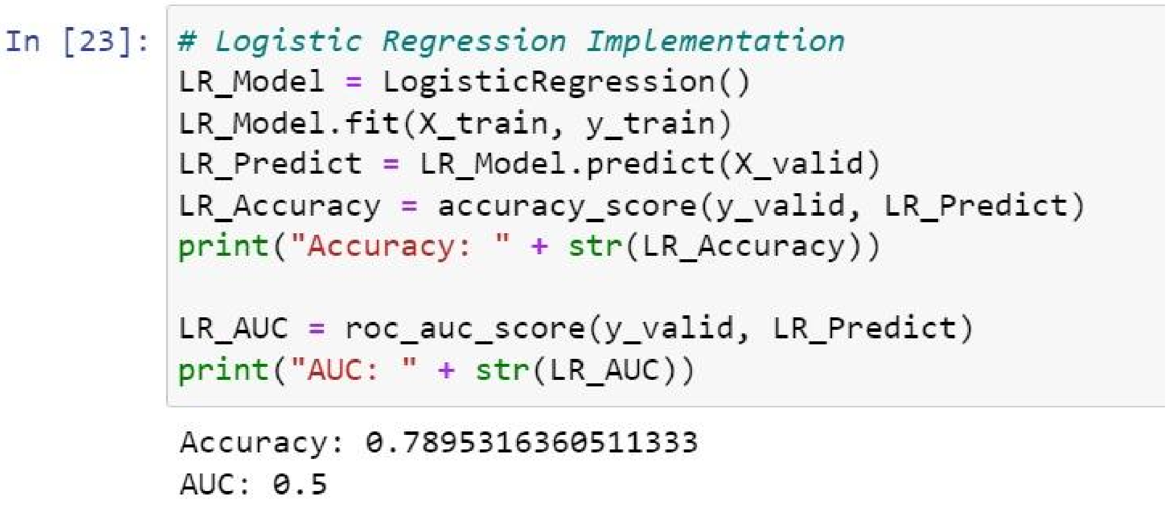


```
In [24]: # Random Forest Implementation
         RFC_Model = RandomForestClassifier()
         RFC_Model.fit(X_train, y_train)
         RFC_Predict = RFC_Model.predict(X_valid)
         RFC_Accuracy = accuracy_score(y_valid, RFC_Predict)
         print("Accuracy: " + str(RFC_Accuracy))

         RFC_AUC = roc_auc_score(y_valid, RFC_Predict)
         print("AUC: " + str(RFC_AUC))

         Accuracy: 0.9216081211730583
         AUC: 0.8338420240978945
```

Figure 1.2
Logistic Regression



```
In [23]: # Logistic Regression Implementation
         LR_Model = LogisticRegression()
         LR_Model.fit(X_train, y_train)
         LR_Predict = LR_Model.predict(X_valid)
         LR_Accuracy = accuracy_score(y_valid, LR_Predict)
         print("Accuracy: " + str(LR_Accuracy))

         LR_AUC = roc_auc_score(y_valid, LR_Predict)
         print("AUC: " + str(LR_AUC))

         Accuracy: 0.7895316360511333
         AUC: 0.5
```

Figure 1.3
Music Loudness vs Energy

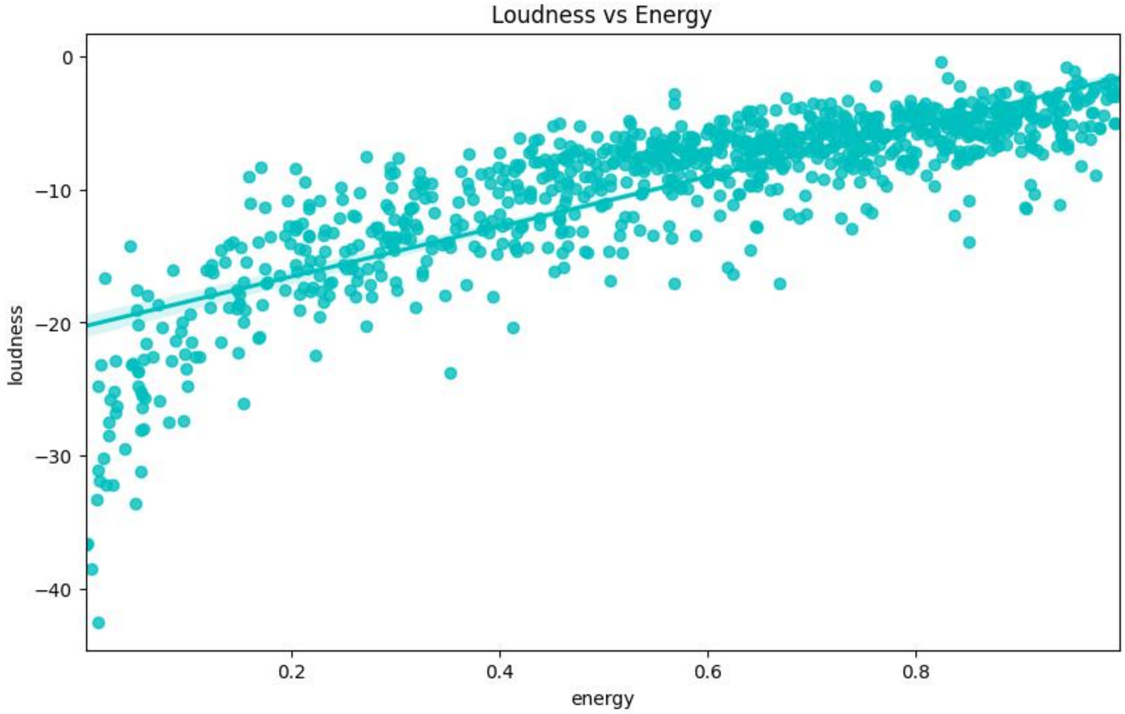


Figure 1.4

Duration of Songs in various Genres

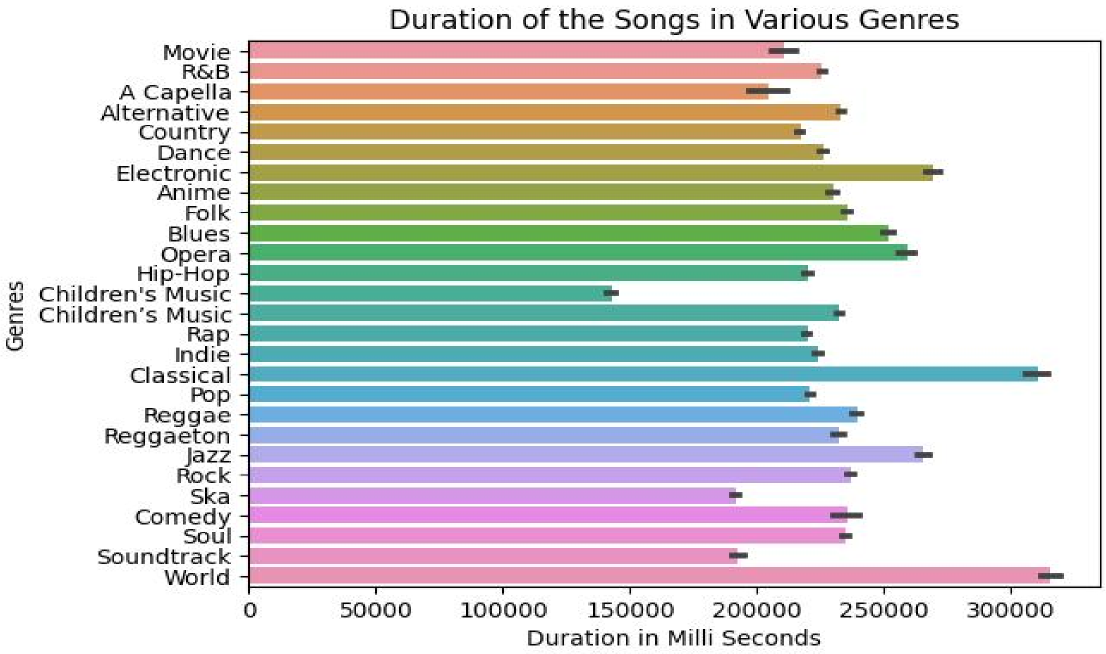
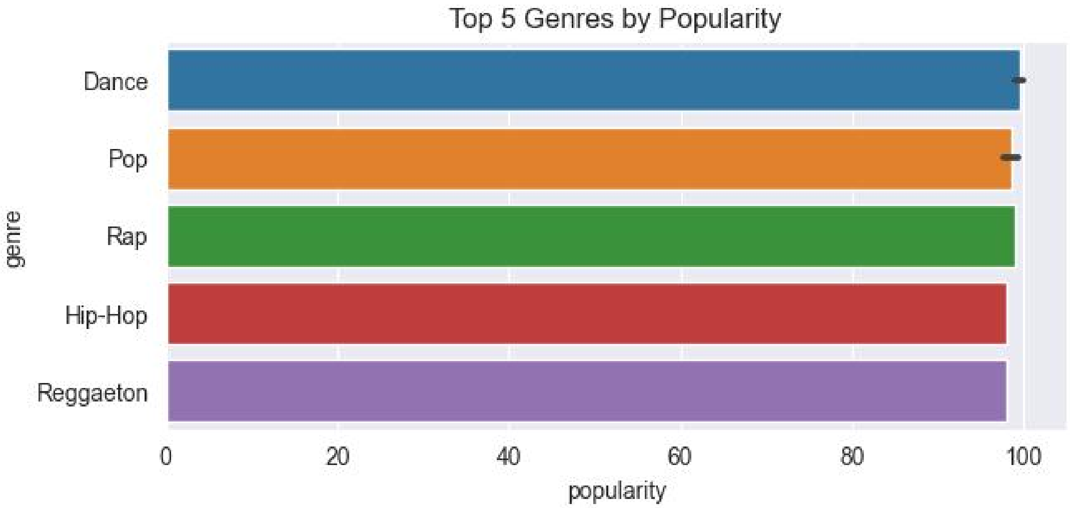


Figure 1.5
Top 5 Genres by Popularity



Last Chapter

5.1 Conclusion

The Spotify data analysis conducted provided valuable insights into music trends and characteristics. Here are some key conclusions drawn from the analysis:

1. Popular vs. Least Popular Songs:

- Identified both the top 10 least popular and most popular songs on Spotify based on the 'popularity' metric.
- This analysis helps understand the range and distribution of popularity among songs in the dataset.

2. Feature Engineering:

- Converted the time duration of songs from milliseconds to seconds, simplifying the time representation for easier analysis.
- Sampled a subset of the data for further analysis, ensuring computational efficiency without compromising insights.

3. Exploratory Data Analysis:

- Visualized the correlation between 'Loudness' and 'Energy' in music, providing insights into the relationship between these two features.
- Explored the relationship between loudness and energy, crucial factors in music composition.

4. Machine Learning Model:

- Developed a Random Forest classifier to predict music genres based on features like 'popularity' and 'duration'.
- Achieved a certain level of accuracy in genre classification, indicating the model's effectiveness in predicting music genres.

5. Genre Analysis:

- Analyzed the duration of songs across different genres, providing an overview of how song lengths vary within each genre.
- Identified and visualized the top 10 genres by popularity, indicating the most popular music genres within the dataset.

Overall, this Spotify data analysis project offered insights into song popularity, feature engineering, correlation analysis, machine learning model training, and genre popularity. These conclusions can be valuable for music enthusiasts, data analysts, and industry professionals interested in understanding music trends and patterns within the Spotify dataset.

5.2 The future scope of Spotify Data Analysis offers exciting potential for further research and developments in the field of music analytics. Here are some prospective areas for development:

1. Personalised Recommendations: Improving recommendation systems by using user listening patterns, preferences, and behaviour to give more tailored music recommendations.

2. Sentiment Analysis: Using sentiment analysis techniques to assess listeners' emotional responses to music, which allows for the production of emotionally tailored playlists.

3. Trend Forecasting: Using predictive analytics to estimate future music trends, artists, record labels, and streaming platforms may remain ahead of the competition.

4. User Segmentation: Using clustering algorithms to group consumers based on their music interests, allowing for targeted marketing techniques and bespoke content distribution.

5. Genre Fusion Analysis: Investigating the creation of new music genres using genre fusion analysis to uncover emerging trends and genre crossovers.

6. Interactive Data Visualisation: Creating interactive dashboards and visualisations that allow users to examine music data interactively, hence increasing user engagement and data understanding.

References

1. Kaggle: Zaheenhamidani. (n.d.). Ultimate Spotify Tracks DB. Retrieved from https://www.kaggle.com/datasets/zaheenhamidani/ultimate-spotify-tracks-db

2. Spotify API Documentation. (n.d.). Retrieved from https://developer.spotify.com/ Information retrieved from Spotify API regarding the approximate number of tracks per genre, totaling 232,725 tracks across 26 genres.

3. Simplilearn. (n.d.). Retrieved from https://shorturl.at/gjqT8

4. Pandas Documentation. (n.d.). Retrieved from https://pandas.pydata.org/Pandas_Cheat_Sheet.pdf

5. scikit-learn. (n.d.). Retrieved from https://scikit-learn.org/stable/user_guide.html

Synopsis

The document is a project report on Spotify Data Analysis submitted by Roopak Mallik in partial fulfillment of the requirements for the Bachelor of Computer Applications (BCA) degree at Manipal University Jaipur. The project's main objective is to provide relevant insights for the music industry by analyzing various aspects of Spotify data, including user listening habits, genre popularity, trends, and artist dynamics. The report outlines the project's organization, conceptual overview, technologies involved, methodology, implementation, results, and conclusions.

The project report is organized into several sections, including an acknowledgment expressing gratitude to faculty members for their support and guidance in the data analysis project. It also includes an abstract outlining the main goal of the data analysis, which is to provide relevant information for the music industry by exploring various aspects of Spotify data. The introduction discusses the impact of streaming services like Spotify on music consumption and the potential insights that can be gained from Spotify's vast collection of music.

The report delves into the conceptual overview of Spotify Data Analysis, including data collection, processing, analysis, insight generation, recommendation engine, performance evaluation, and continuous improvement. It also covers the methodology, which involves data import, initial exploration, data preparation, feature engineering, exploratory data analysis, and machine learning model development.

The results and analysis section presents key conclusions drawn from the analysis, including insights into popular vs. least popular songs, feature engineering, correlation analysis, machine learning model training, and genre popularity. Additionally, the report discusses the use of visualizations and charts to shed light on the complex network of relationships between musicians, songs, genres, and listeners within the Spotify music ecosystem.