Statistical Modelling using Python

Understanding and Analyzing Music Features for Popularity Prediction

Kamal Nadh Lukka
13/09/2024

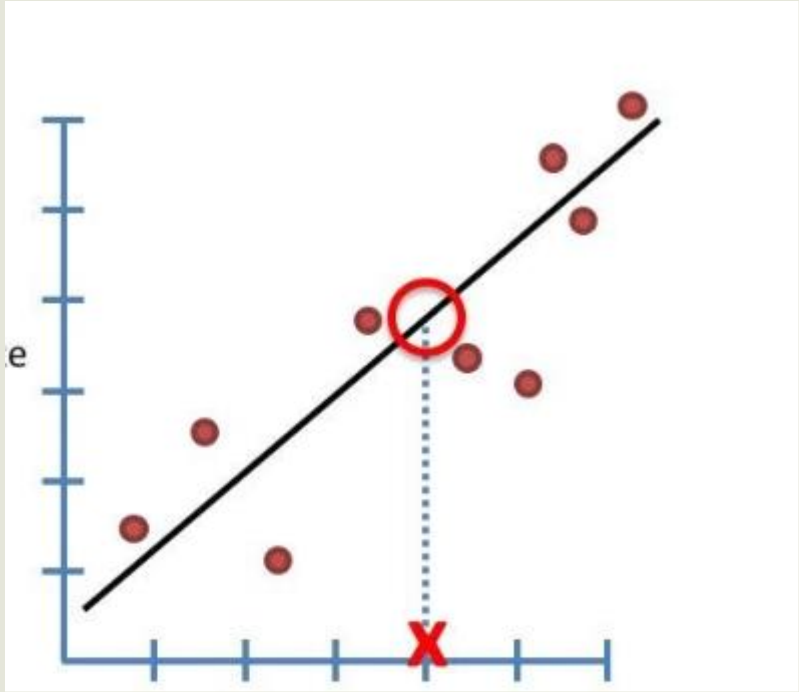
INTRODUCTION

What is Statistical Modelling?

- A mathematical framework to estimate or predict real-world behaviors.
- Describes relationships between variables through equations.

Focus:

- Analyze how music features affect popularity.



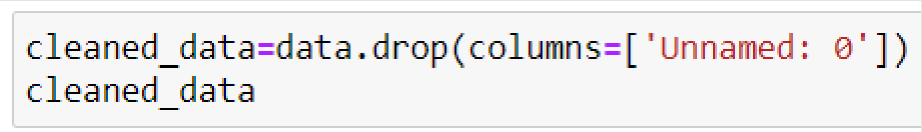
DATASET OVERVIEW

Music Dataset:

- Track Name, Artists, Album Name, Popularity, Duration, Energy, Danceability, etc.
- 22 columns including both numerical and categorical features.

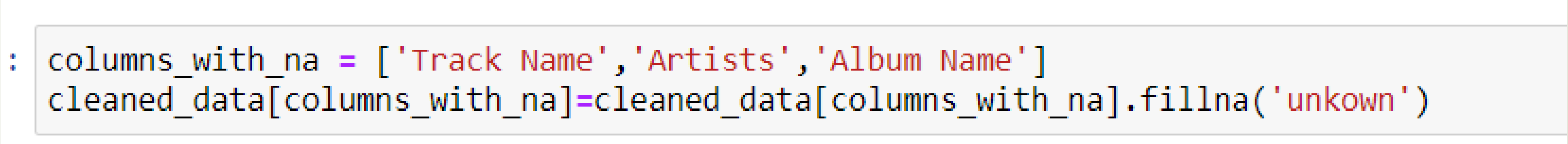
DATA CLEANING

- **Steps Taken:**
  - Removed unnecessary 'Unnamed: 0' column.



```
cleaned_data=data.drop(columns=['Unnamed: 0'])
cleaned_data
```

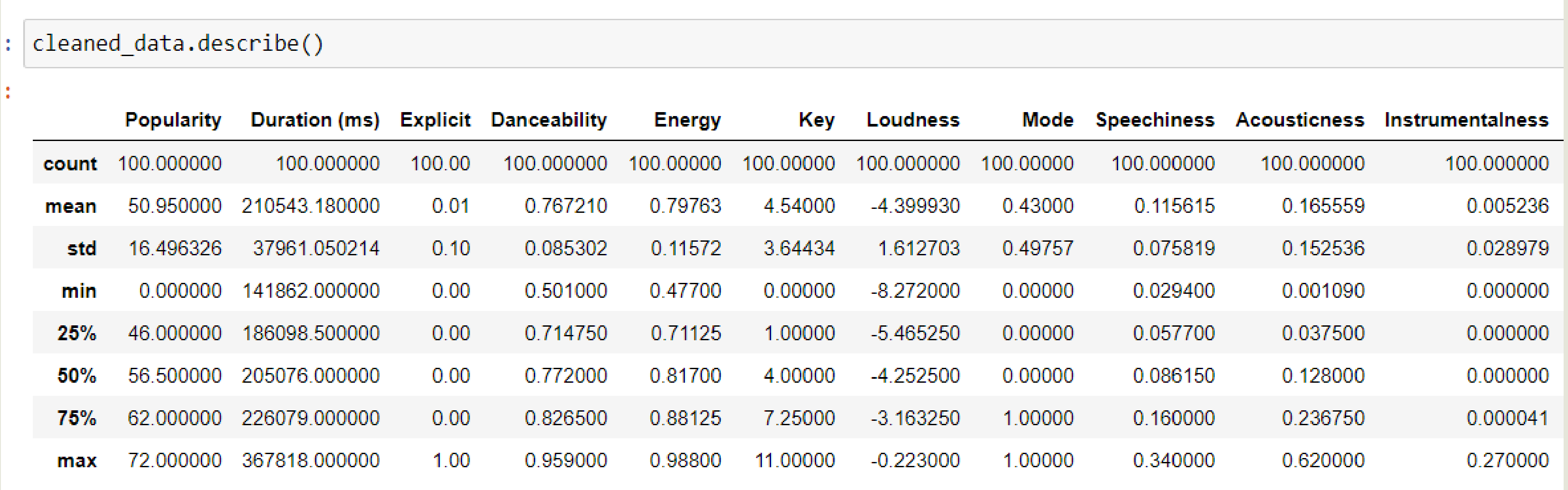
  - Handled missing values by filling in missing Track Name, Artists, and Album Name with "Unknown."



```
columns_with_na = ['Track Name','Artists','Album Name']
cleaned_data[columns_with_na]=cleaned_data[columns_with_na].fillna('unkown')
```

Initial Data Insights

- **Statistical Summary:**
  - **Average Popularity:** 50.95
  - **Energy Mean:** 0.798
  - **Danceability Mean:** 0.767
  - **Key Insights:**
    - Tracks vary widely in terms of energy and danceability.

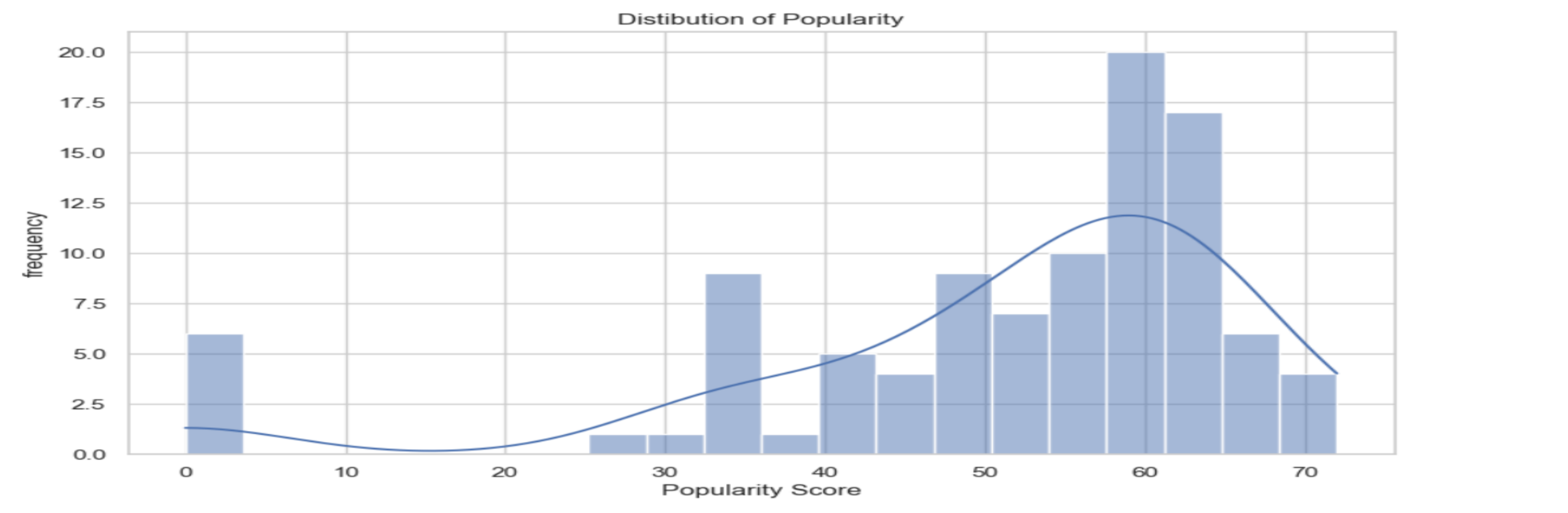


```
cleaned_data.describe()
```

	Popularity	Duration (ms)	Explicit	Danceability	Energy	Key	Loudness	Mode	Speechiness	Acousticness	Instrumentalness
count	100.000000	100.000000	100.00	100.000000	100.00000	100.00000	100.000000	100.00000	100.000000	100.000000	100.000000
mean	50.950000	210543.180000	0.01	0.767210	0.79763	4.54000	-4.399930	0.43000	0.115615	0.165559	0.005236
std	16.496326	37961.050214	0.10	0.085302	0.11572	3.64434	1.612703	0.49757	0.075819	0.152536	0.028979
min	0.000000	141862.000000	0.00	0.501000	0.47700	0.00000	-8.272000	0.00000	0.029400	0.001090	0.000000
25%	46.000000	186098.500000	0.00	0.714750	0.71125	1.00000	-5.465250	0.00000	0.057700	0.037500	0.000000
50%	56.500000	205076.000000	0.00	0.772000	0.81700	4.00000	-4.252500	0.00000	0.086150	0.128000	0.000000
75%	62.000000	226079.000000	0.00	0.826500	0.88125	7.25000	-3.163250	1.00000	0.160000	0.236750	0.000041
max	72.000000	367818.000000	1.00	0.959000	0.98800	11.00000	-0.223000	1.00000	0.340000	0.620000	0.270000

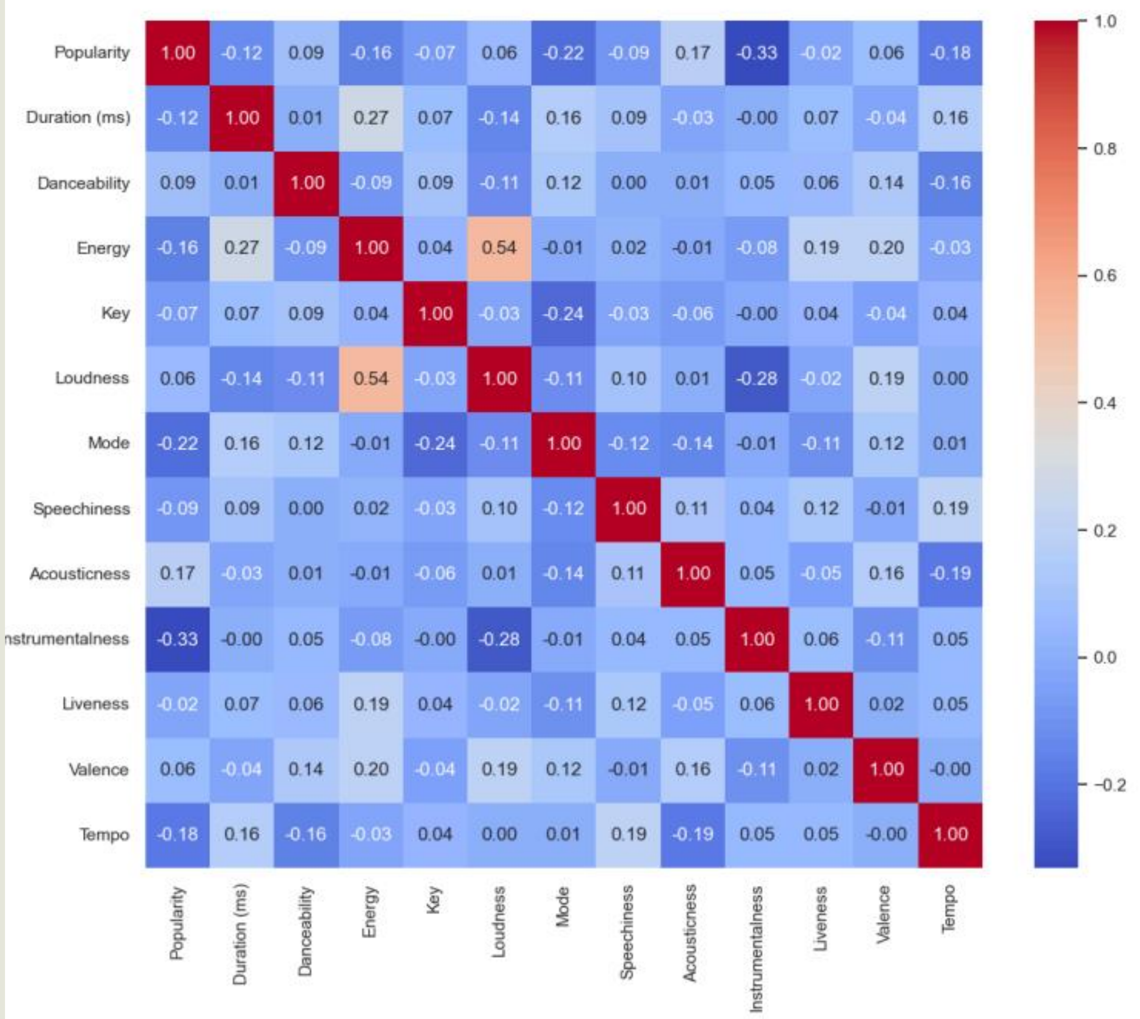
Distribution of Popularity

- **Popularity Distribution Visualization:**
- Most tracks have a popularity score between 40 to 70.
- Peaks around 50 and 60, indicating moderate to high popularity.



Correlation Matrix

- **Heatmap Insights:** Positive correlations between:
    - Popularity & Energy
    - Popularity & Loudness
- Negative correlation between Popularity & Acousticness.



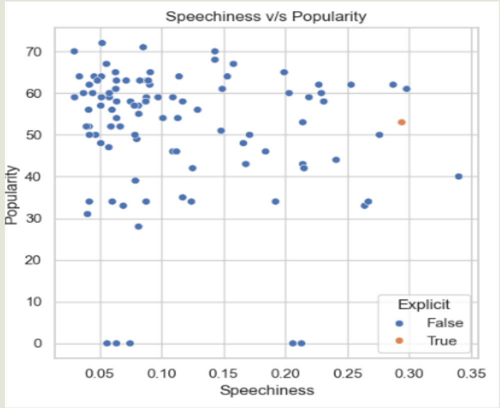
	Popularity	Duration (ms)	Danceability	Energy	Key	Loudness	Mode	Speechiness	Acousticness	Instrumentalness	Liveness	Valence	Tempo
Popularity	1.00	-0.12	0.09	-0.16	-0.07	0.06	-0.22	-0.09	0.17	-0.33	-0.02	0.06	-0.18
Duration (ms)	-0.12	1.00	0.01	0.27	0.07	-0.14	0.16	0.09	-0.03	-0.00	0.07	-0.04	0.16
Danceability	0.09	0.01	1.00	-0.09	0.09	-0.11	0.12	0.00	0.01	0.05	0.06	0.14	-0.16
Energy	-0.16	0.27	-0.09	1.00	0.04	0.54	-0.01	0.02	-0.01	-0.08	0.19	0.20	-0.03
Key	-0.07	0.07	0.09	0.04	1.00	-0.03	-0.24	-0.03	-0.06	-0.00	0.04	-0.04	0.04
Loudness	0.06	-0.14	-0.11	0.54	-0.03	1.00	-0.11	0.10	0.01	-0.28	-0.02	0.19	0.00
Mode	-0.22	0.16	0.12	-0.01	-0.24	-0.11	1.00	-0.12	-0.14	-0.01	-0.11	0.12	0.01
Speechiness	-0.09	0.09	0.00	0.02	-0.03	0.10	-0.12	1.00	0.11	0.04	0.12	-0.01	0.19
Acousticness	0.17	-0.03	0.01	-0.01	-0.06	0.01	-0.14	0.11	1.00	0.05	-0.05	0.16	-0.19
nstrumentalness	-0.33	-0.00	0.05	-0.08	-0.00	-0.28	-0.01	0.04	0.05	1.00	0.06	-0.11	0.05
Liveness	-0.02	0.07	0.06	0.19	0.04	-0.02	-0.11	0.12	-0.05	0.06	1.00	0.02	0.05
Valence	0.06	-0.04	0.14	0.20	-0.04	0.19	0.12	-0.01	0.16	-0.11	0.02	1.00	-0.00
Tempo	-0.18	0.16	-0.16	-0.03	0.04	0.00	0.01	0.19	-0.19	0.05	0.05	-0.00	1.00

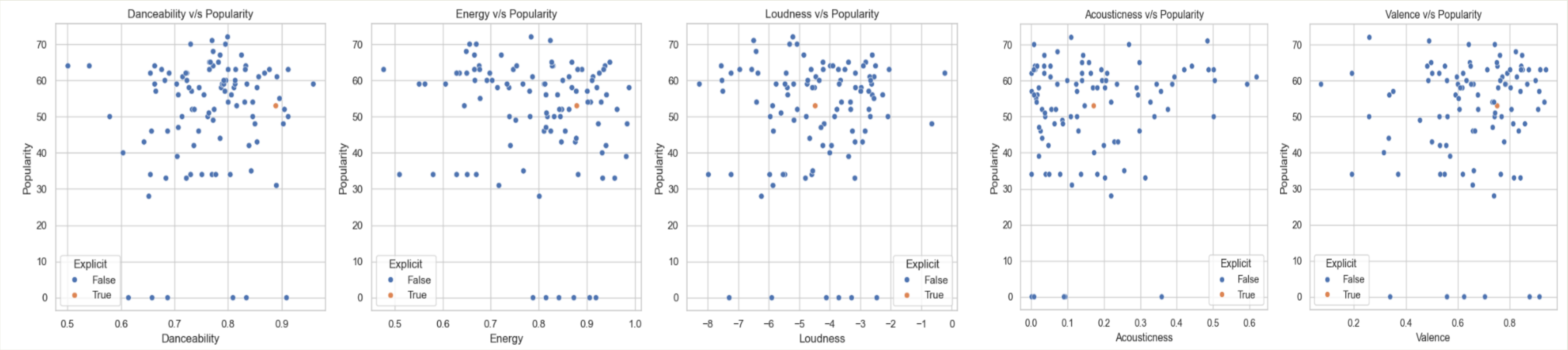
Feature Analysis

- **Key Features Influencing Popularity:**
- **Danceability:** Tracks that are easier to dance to are slightly more popular.
- **Energy & Loudness:** Higher energy and louder tracks are more popular.
- **Acousticness:** Tracks with more acoustic elements tend to be less popular.
- **Valence:** Happier-sounding tracks tend to be more popular.

Feature vs Popularity Plots

- **Visualization of Feature Impact on Popularity:**
- **Danceability vs Popularity** **Energy vs Popularity**
- **Loudness vs Popularity** **Acousticness vs Popularity**
- **Valence vs Popularity**





Statistical Modelling Approach

- **Model Used:**
  - Linear Regression to model relationships between features like energy, loudness, and popularity.
  - Model aimed at quantifying how these features predict popularity.

Conclusion

- **Key Takeaways:**
  - Popularity is influenced by energy, loudness, and danceability.
  - Statistical modelling can provide insight into what makes music tracks popular.
  - Future work can explore more complex models or focus on improving predictive accuracy.