Spotify Data Analysis & Popularity Prediction

Gahyoung Lee, Yu-Chiao Shaw, Yichun Jin,
Chia-Yun Chiang, Yuwei Hsu

Professor Fidel Rodriguez
Northeastern University

INDEX

OVERVIEW

Has the music streaming service influenced the creation of a song? We would like to use the Spotify dataset to understand how the song is composed and what feature is affecting the popularity of the song. We will analyze multiple features of music, such as energy, acousticness, danceability, valence, explicit, and popularity. It can provide us with a deep understanding of the spectrum of elements and the trend of music across a wide period. Then, we will build predictive models using machine learning and data mining; one for predicting the explicit and another for predicting the popularity of the song. The model can be applied as a popularity predictor for music producers, songwriters, marketers, streaming companies, etc.

Research Questions

1. Explore the trend in music with various features.
2. Recommend the top three popular artists based on the specific music feature.
3. Find any characteristics related to the explicit songs.
4. Predict the popularity based on different features of songs.

Dataset Description 01

I. Trend Analysis Scope

The original data set consists of 19 attributes with 174,389 rows of Spotify music from 1920 to 2021.

II. Prediction Analysis Scope

The filtered dataset consists of 19 attributes with 42,371 rows of Spotify music from 2000 to 2021.

Features description

Numerical Features:

- **acousticness:** A confidence measure from 0.0 to 1.0 of whether the track is acoustic. (1.0 represents high confidence the track is acoustic.)
- **danceability:** Describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. (1.0 is most danceable.)
- **duration_ms:** The duration of the track in milliseconds.
- **energy:** Represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy.
- **id:** The Spotify ID for the track.
- **instrumentalness:** Predicts whether a track contains no vocals. (1.0 indicates the greater likelihood the track contains no vocal content.

Dataset Description 02

Features description

Numerical Features:

- **key:** The estimated overall key of the track. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1.
- **liveness:** Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live.
- **loudness:** Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Values typical range between -60 and 0 decibels (db).
- **release_date**
- **speechiness:** Detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value.
- **tempo:** The overall estimated tempo of a track in beats per minute (BPM). Values typically range between 50 and 150.
- **valence:** A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive, while tracks with low valence sound more negative.
- **year**

Dataset Description 03

Features description

Target Feature:

- **popularity:** Song ratings of Spotify audience.

Dummy Features:

- **explicit:** Explicit = 1 track is one that has curse words or language or art that is sexual, violent, or offensive in nature.
- **mode:** Indicates the modality (major = 1 or minor = 0) of a track.

Categorical Features:

- **artists:** Artists of the tracks.
- **name:** Name of the songs.

Dataset Description 04

Data Cleaning

1. Drop extraneous columns and NaN values.
2. Find duplicate rows with the same artist and song name and delete them.
3. Drop rows with 0 tempo (tempo = 0 is not reasonable. Tempo typically ranges from 50 to 150).
4. Change songs' duration from milliseconds to minutes.

After cleaning, the data set has 37,809 rows with 15 columns.

Exploratory data analysis

01

I. Numerical data description

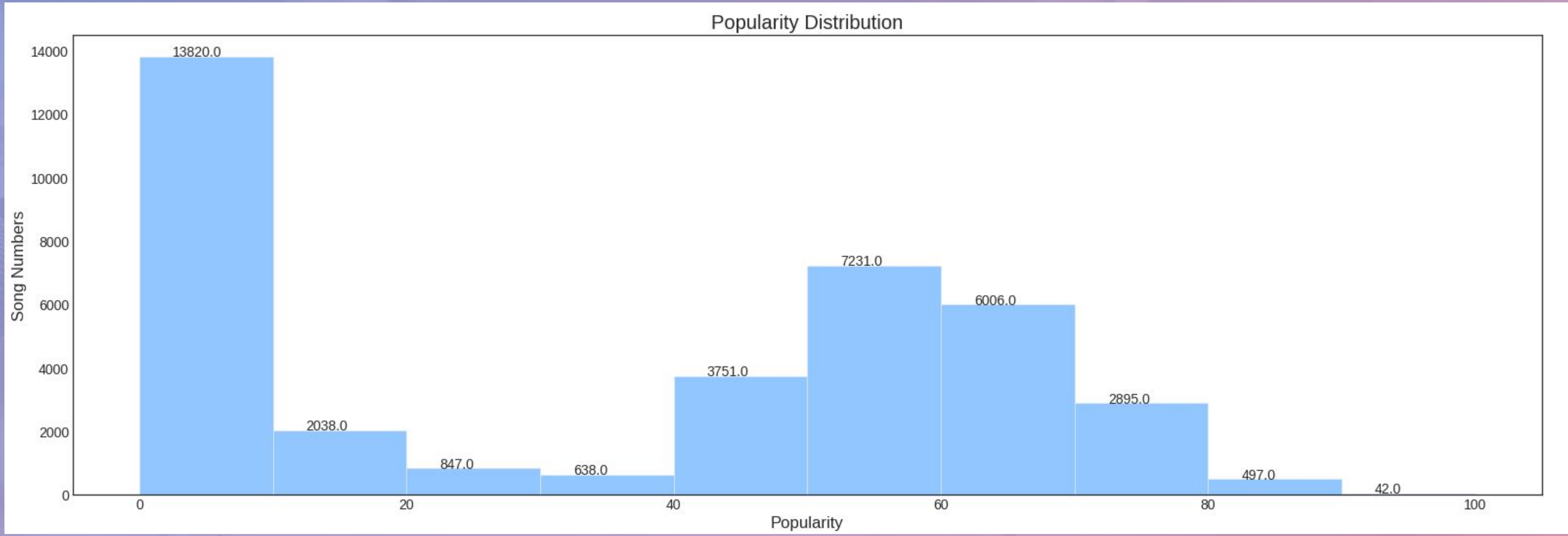
	acousticness	danceability	energy	explicit	instrumentalness	key	liveness	loudness	mode	popularity	speechiness	tempo	valence	year	duration_mins
count	37765.00	37765.00	37765.00	37765.00	37765.00	37765.00	37765.00	37765.00	37765.00	37765.00	37765.00	37765.00	37765.00	37765.00	37765.00
mean	0.25	0.58	0.66	0.16	0.20	5.32	0.22	-8.10	0.65	34.45	0.09	122.87	0.48	2011.93	4.07
std	0.31	0.17	0.24	0.37	0.34	3.57	0.19	4.57	0.48	28.49	0.11	28.16	0.26	6.13	3.15
min	0.00	0.06	0.00	0.00	0.00	0.00	0.01	-45.35	0.00	0.00	0.02	34.63	0.00	2000.00	0.28
25%	0.01	0.47	0.50	0.00	0.00	2.00	0.10	-9.60	0.00	1.00	0.04	100.12	0.27	2007.00	3.08
50%	0.09	0.59	0.69	0.00	0.00	6.00	0.13	-7.00	1.00	46.00	0.05	125.00	0.47	2013.00	3.72
75%	0.42	0.71	0.86	0.00	0.25	8.00	0.29	-5.21	1.00	59.00	0.10	139.02	0.68	2017.00	4.51
max	1.00	0.99	1.00	1.00	1.00	11.00	1.00	3.37	1.00	100.00	0.96	221.95	1.00	2021.00	88.97

- The mean popularity score of the data set is 34.45. 50% of the songs have popularity under 46 and 75% of the songs have popularity under 59.
- The mean duration of songs is 4.07, 50% of the songs are with 3.72 minutes.
- Only 16% of the songs are explicit.

Exploratory data analysis

02

II. Data Visualization

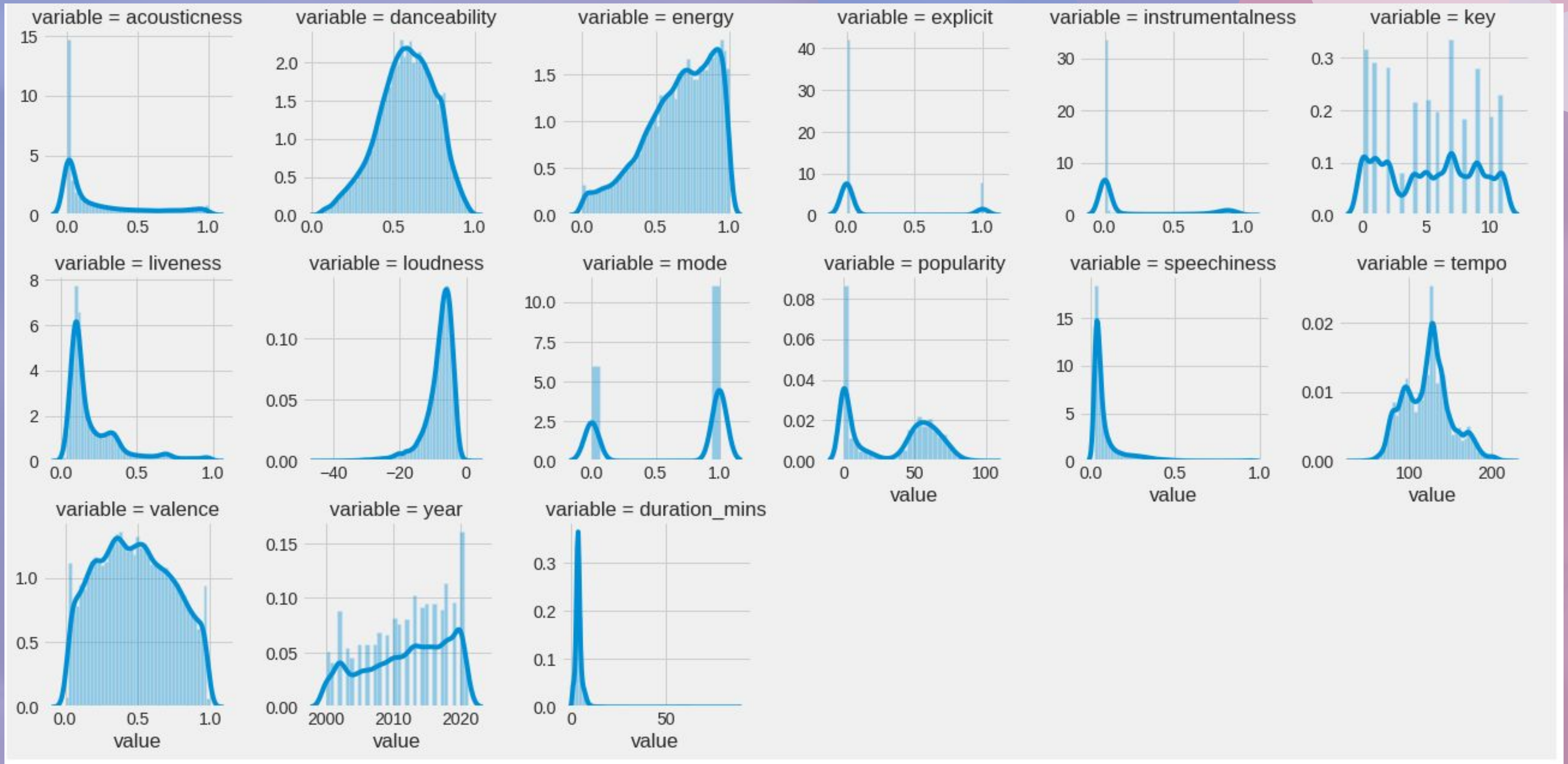


- 36.6% of the songs have popularity score of 0.
- Popularity ranges 40 to 80 have the most songs.
- Only 579 (1.5%) songs have popularity above 80.

Exploratory data analysis

03

II. Data Visualization - Distplot by features

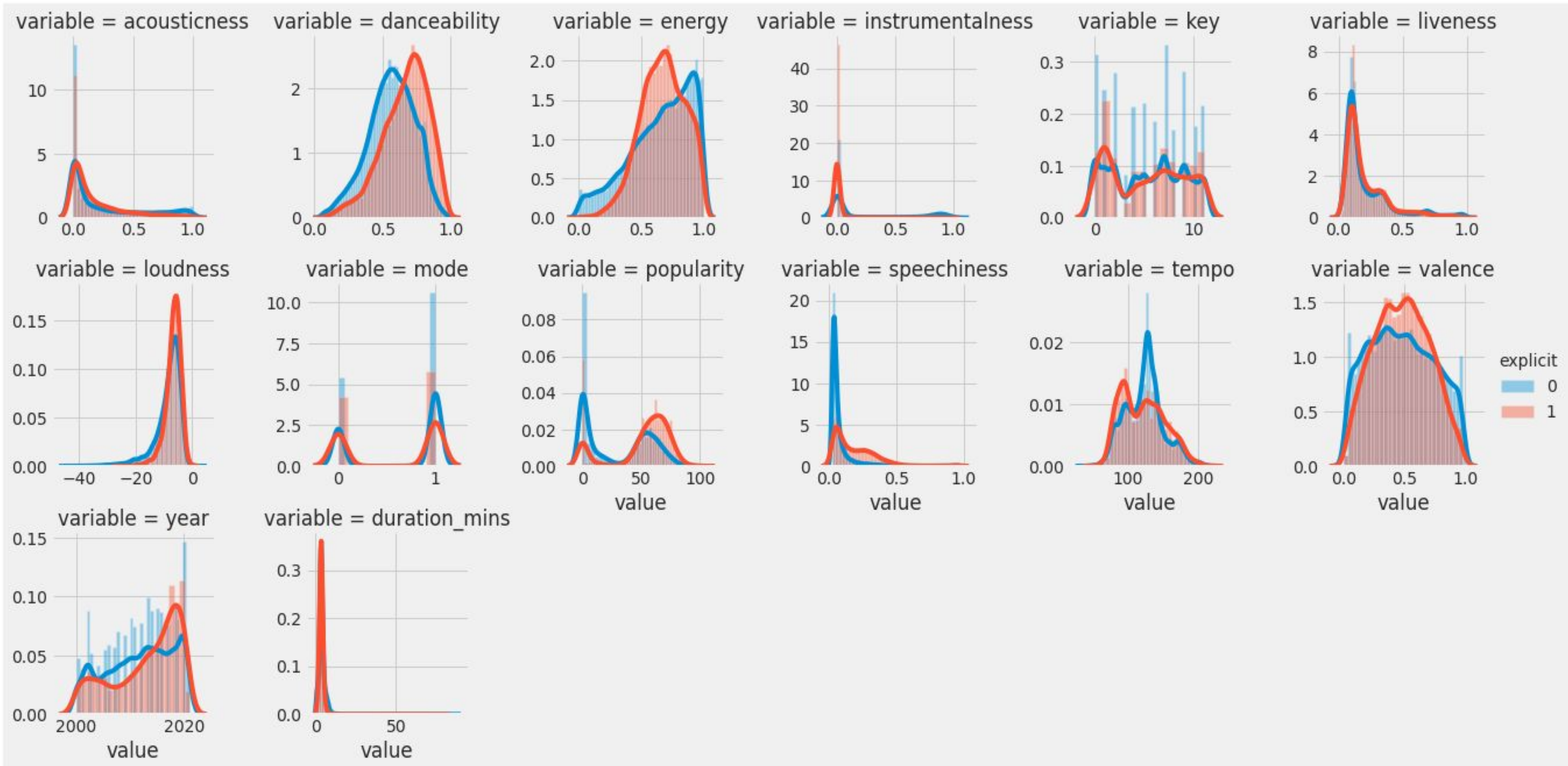


- Acounticness, liveness, and speechiness have a left-skewed distribution. Most songs have low values in these features.
- Energy and loudness show a right-skewed distribution. It means most songs have high energy and loudness.
- Danceability, tempo, and valence are normally distributed.
- The total number of songs increased year by year.

Exploratory data analysis

04

II. Data Visualization - Distplot by explicit

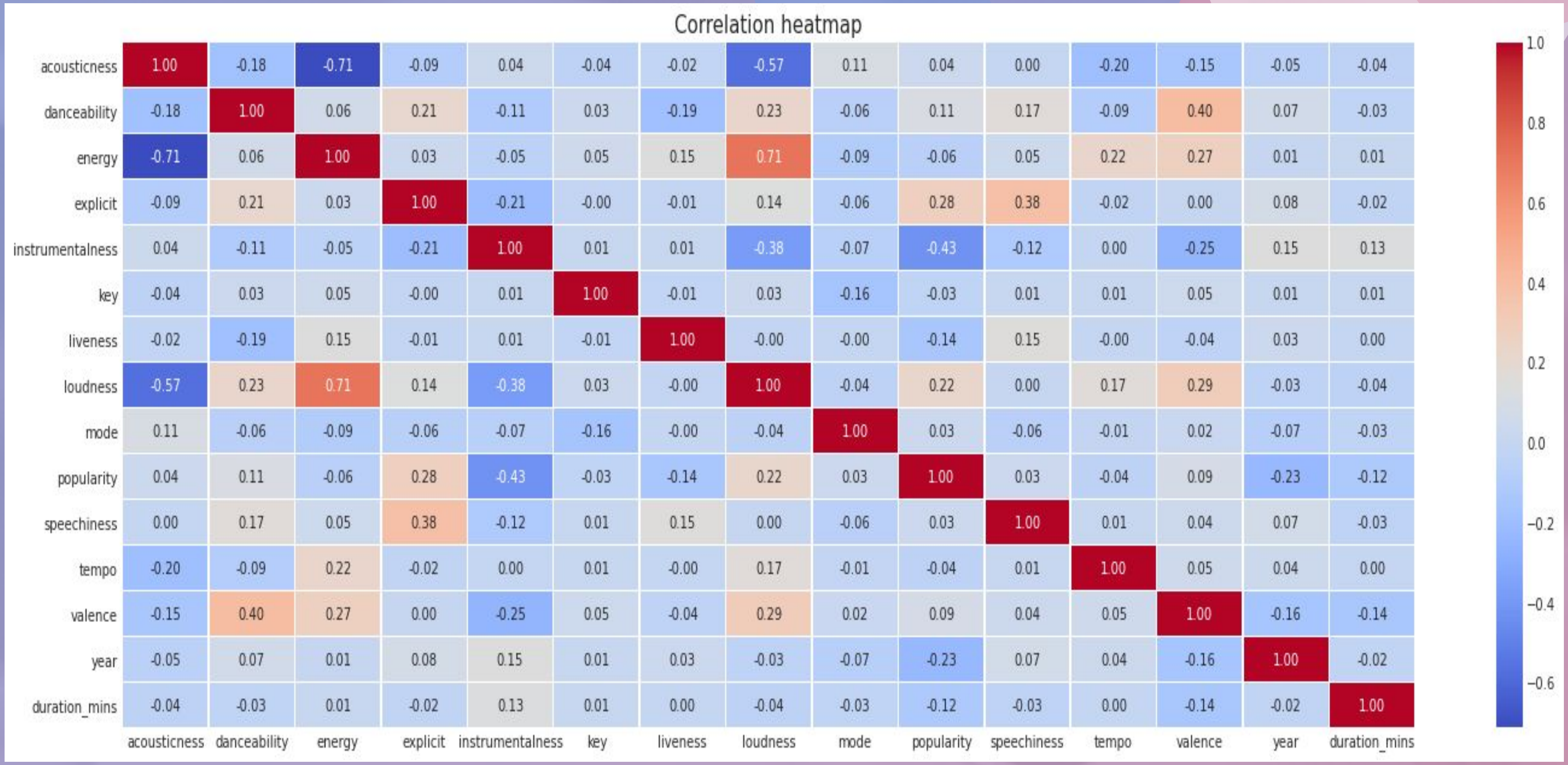


- Songs containing explicit contents are moderately danceable, energetic, and have higher valence.
- Explicit songs tend to have higher popularity scores, and these songs have been released more in recent years.

Exploratory data analysis

05

II. Correlation between features



- The popularity may have a positive relationship with explicit, loudness, and a negative relationship to instrumentalness and year.
- Loudness has a strong positive relationship with energy and negative relationships with acousticness and instrumentalness. This makes sense as energy is influenced by the volume of the song being played.
- Explicit may be related to speechiness and danceability.

Exploratory data analysis

06

II. Artists recommendation

- We created a function for artists recommendation with different input, here are samples for a few different inputs:

	artists	acousticness	danceability	duration_ms	energy	explicit	instrumentalness	key	liveness	loudness	mode	popularity	speechiness	tempo	valence	year
242	['24kGoldn', 'iann dior']	0.2210	0.700	140526.0	0.722	1.0	0.000000	7.0	0.2720	-3.558	0.0	96.0	0.0369	90.989	0.756	2020.0
1660	['Bad Bunny', 'Jhay Cortez']	0.4010	0.731	205090.0	0.573	1.0	0.000052	4.0	0.1130	-10.059	0.0	95.0	0.0544	109.928	0.145	2020.0
1665	['Bad Bunny', 'ROSALÃ A']	0.0303	0.856	203201.0	0.618	0.0	0.000000	7.0	0.0866	-4.892	1.0	94.0	0.2860	81.993	0.391	2020.0

1) **Top 3 popular artists:** ['24kGoldn', 'iann dior'], ['Bad Bunny', 'Jhay Cortez'], ['Bad Bunny', 'ROSALÃ A']

	artists	acousticness	danceability	duration_ms	energy	explicit	instrumentalness	key	liveness	loudness	mode	popularity	speechiness	tempo	valence	year
1665	['Bad Bunny', 'ROSALÃ A']	0.0303	0.856	203201.0	0.618	0.0	0.0	7.0	0.0866	-4.892	1.0	94.0	0.286	81.993	0.391	2020.0
2594	['Cardi B', 'Megan Thee Stallion']	0.0194	0.935	187541.0	0.454	1.0	0.0	1.0	0.0824	-7.509	1.0	92.0	0.375	133.073	0.357	2020.0
11563	['Pop Smoke', 'Lil Baby', 'DaBaby']	0.1140	0.823	190476.0	0.586	1.0	0.0	6.0	0.1930	-6.606	0.0	91.0	0.200	125.971	0.347	2020.0

2) **Top 3 popular artists with high danceability:** ['Bad Bunny', 'ROSALÃ A'], ['Cardi B', 'Megan Thee Stallion'], ['Pop Smoke', 'Lil Baby', 'DaBaby']

	artists	acousticness	danceability	duration_ms	energy	explicit	instrumentalness	key	liveness	loudness	mode	popularity	speechiness	tempo	valence	year
4904	['Erik Eriksson', 'White Noise Baby Sleep', 'W...	0.7910	0.000	90228.0	0.00342	0.0	1.000	8.0	0.1110	-28.460	1.0	84.0	0.0000	0.000	0.0000	2017.0
9012	['Ludovico Einaudi', 'Daniel Hope', 'I Virtuos...	0.9340	0.447	315427.0	0.44900	0.0	0.961	2.0	0.0697	-10.634	1.0	76.0	0.0376	92.468	0.0360	2013.0
12395	['RÅ®DE']	0.0347	0.655	205766.0	0.48700	0.0	0.897	6.0	0.2710	-7.988	1.0	76.0	0.0330	139.914	0.0454	2017.0

3) **Top 3 popular artists with high instrumentalness:** ['Erik Eriksson', 'White Noise Baby Sleep', 'White Noise for Babies'], ['Ludovico Einaudi', 'Daniel Hope', 'I Virtuosi Italiani'], ['RÅ®DE']

Exploratory data analysis

07

II. Music features recommendation

- We observe the music features of the tracks with popularity above 75 and calculate a 95% confidence interval that the music features are within the range. It could be used as suggestions for artists and composers (eg. Taylor Swift) to reference these features ranges.

	acousticness	danceability	energy	explicit	instrumentalness	key	liveness	loudness	mode	popularity	speechiness	tempo	valence	year	duration_mins
count	1274.00	1274.00	1274.00	1274.00	1274.00	1274.00	1274.00	1274.00	1274.00	1274.00	1274.00	1274.00	1274.00	1274.00	1274.00
mean	0.25	0.67	0.63	0.41	0.01	5.33	0.17	-6.39	0.60	79.86	0.11	120.87	0.48	2017.20	3.41
std	0.26	0.14	0.18	0.49	0.09	3.58	0.12	2.52	0.49	3.90	0.10	29.34	0.22	4.24	0.74
min	0.00	0.21	0.02	0.00	0.00	0.00	0.03	-24.64	0.00	76.00	0.02	48.72	0.04	2000.00	1.08
25%	0.04	0.58	0.52	0.00	0.00	2.00	0.10	-7.67	0.00	77.00	0.04	97.01	0.31	2017.00	2.95
50%	0.15	0.68	0.64	0.00	0.00	5.00	0.12	-5.95	1.00	79.00	0.06	120.00	0.47	2019.00	3.38
75%	0.39	0.77	0.75	1.00	0.00	8.00	0.20	-4.75	1.00	82.00	0.13	140.06	0.64	2020.00	3.79
max	0.98	0.98	0.99	1.00	0.96	11.00	0.85	-1.19	1.00	100.00	0.78	220.10	0.97	2021.00	8.07

- The above chart shows features for songs with popularity above 75.

	min	mean	max
acousticness	0.24	0.25	0.26
danceability	0.66	0.67	0.68
energy	0.62	0.63	0.64
explicit	0.38	0.41	0.44
instrumentalness	0.01	0.01	0.01
key	5.13	5.33	5.53
liveness	0.16	0.17	0.18
loudness	-6.53	-6.39	-6.25
mode	0.57	0.60	0.63
popularity	79.65	79.86	80.07
speechiness	0.10	0.11	0.12
tempo	119.26	120.87	122.48
valence	0.47	0.48	0.49
year	2016.97	2017.20	2017.43
duration_mins	3.37	3.41	3.45

- The chart on the left shows 95% confidence intervals for each feature.

Methodology

We conduct three main methodologies in this project as below:

1. **Trend Analysis**
2. **Hypothesis Testing** (ANOVA and Chi-square independent test)
3. **Data Mining Techniques in Machine Learning**
   (Linear Regression, Logistic Regression, Decision Tree Regressor,
   XGB Regressor, Random Forest Classifier, and K Nearest Neighbors Classifier)

Methodology

- **Trend Analysis**
  (1) Observe trends and seasonality of different features in tracks from 1920 to 2021.
  (2) Use the Augmented Dickey Fuller test to check whether the data is stationary or not, then use differencing methods to transform the dataset.

- **Hypothesis Testing**
  (1) Use ANOVA to test whether explicit (0 or 1) is related to danceability and valence.
  (2) Use Chi-square independent test to test whether explicit (0 or 1) is related to the key.

- **Data Mining Techniques**
  (1) Use a logistic regression model to predict explicit (0 or 1) using all the other music attributes.
  (2) Use different kinds of regression models to predict popularity value (linear regression, decision tree regressor, XGB regressor).
  (3) Use different kinds of classification models to predict popularity level (Random Forest Classifier, logistic regression (multiclass), and K Nearest Neighbors Classifier).

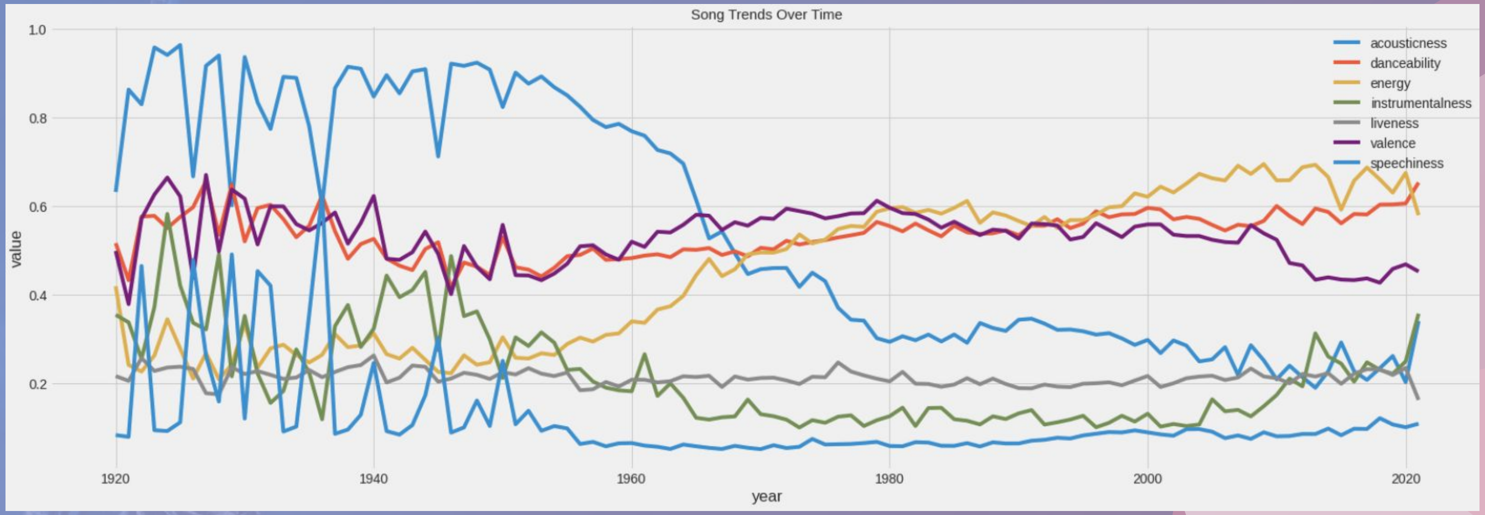
Trend Analysis

1. Observe trends and seasonality of different features in tracks from 1920 to 2021.
2. Use the Augmented Dickey Fuller test to check whether the data is stationary or not, then use differencing methods to transform the dataset.

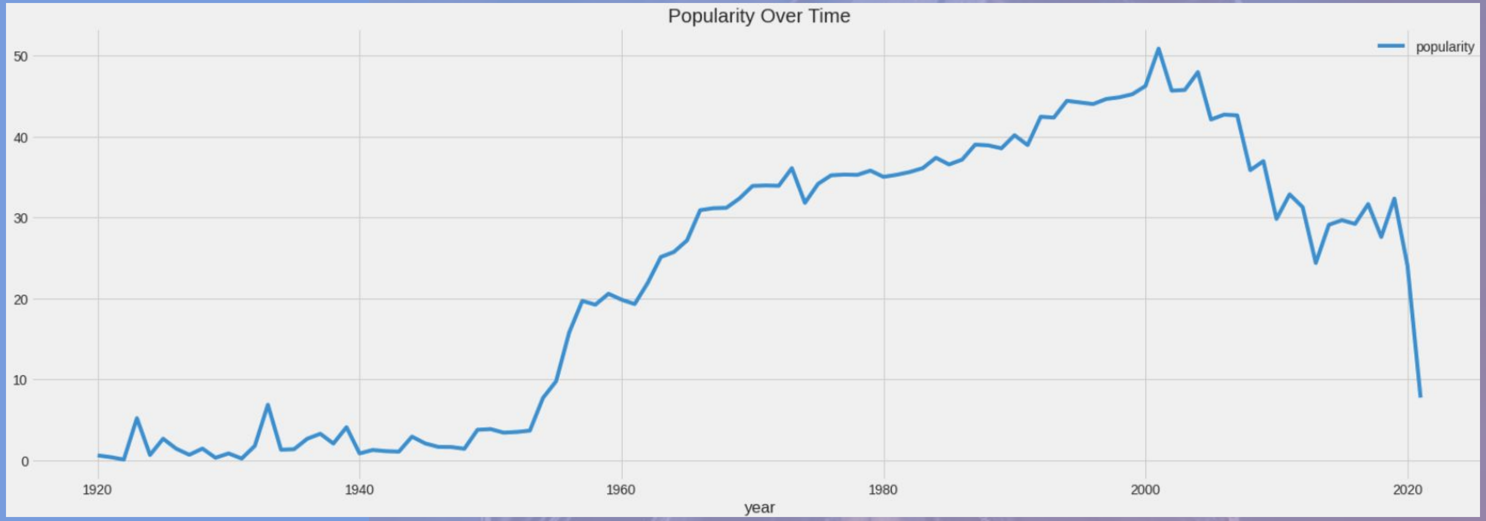
Trend Analysis

01

1) Audio Features:



- Acousticness is decreasing and energy is increasing significantly over years. The possible reason is the rise of development in pop music during the 1950s and 1960s, which is louder, less acoustic and more energetic.
- Speechness has significant changes from 1920 to 1950, and becomes an almost flattened line after 1950.



2) Popularity over time:

- The popularity of the tracks increased from 1950 to 2000 and drastically decreased after the year 2000. The possible reason is that it’s not easy for everyone to adapt to new music.

Trend Analysis

02

3) ADF test of original data:

```
Results of Dickey-Fuller Test:
Test Statistic                 -1.457798
p-value                         0.554275
#Lags Used                      8.000000
Number of Observations Used    93.000000
Critical Value (1%)            -3.502705
Critical Value (5%)            -2.893158
Critical Value (10%)           -2.583637
dtype: float64
```

- H0 = The time series data is not stationary (includes some time-dependent structure).
- H1 = The data is stationary.
- P-value = 0.554
- Test statistics = -1.458
- Test statistic value is still higher than all levels of significance after log transformation. We fail to reject the null hypothesis, hence this is a non-stationary dataset.

4) ADF test of transformed data:

```
Results of Dickey-Fuller Test:
Test Statistic                 -1.034915
p-value                         0.740318
#Lags Used                      7.000000
Number of Observations Used    93.000000
Critical Value (1%)            -3.502705
Critical Value (5%)            -2.893158
Critical Value (10%)           -2.583637
dtype: float64
```

- P-value = 0.74
- Test statistics = -1.034
- The test statistic value is still higher than all levels of significance after log transformation. We fail to reject the null hypothesis, hence the data is non-stationary.
- This dataset is not suitable for time series analysis.

Hypothesis Testing

1. Use ANOVA to test whether explicit (0 or 1) is related to danceability and valence.
2. Use Chi-square independent test to test whether explicit (0 or 1) is related to the key.

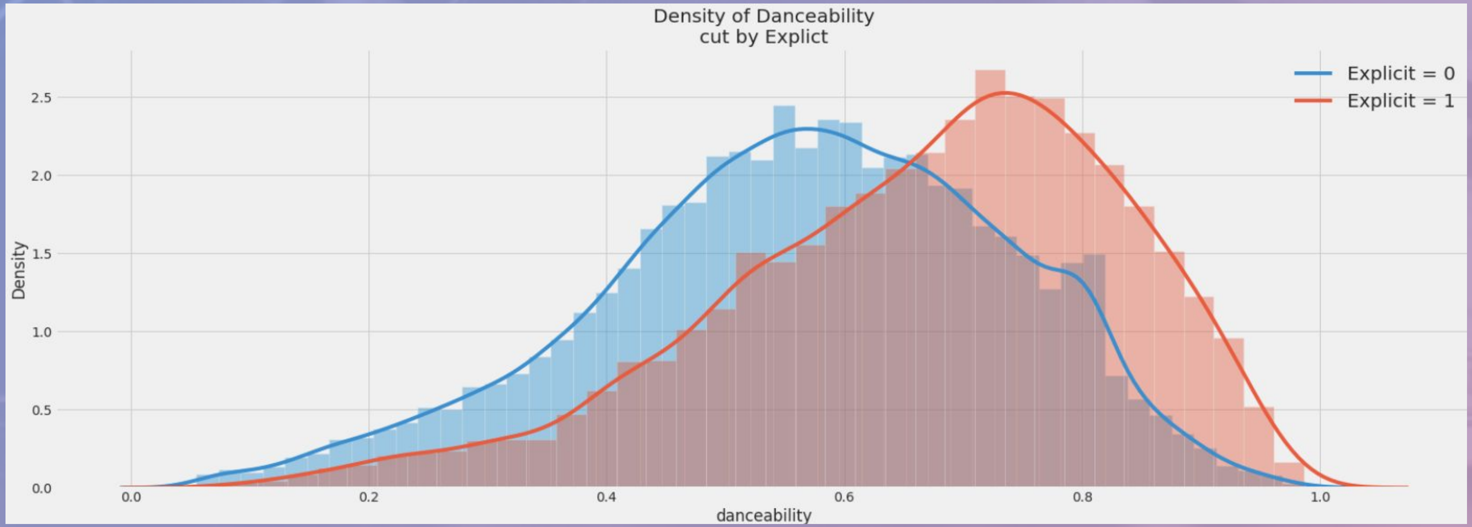
Hypothesis Testing

01

1. Hypothesis test Analysis in danceability

(1) Test Setup

- **The Null Hypothesis (H0):** The danceability of the tracks without explicit lyrics is the same as those songs with explicit content.
- **The Alternative Hypothesis (H1):** The danceability of the tracks without explicit lyrics is different than those songs with explicit songs.
- The alpha-value will be set to 0.05.



(2) Test Analysis

- The claim is H0.
- Since p-value = 0 and it is less than the alpha-value = 0.05. We **reject** the null hypothesis.
- There is enough evidence to reject the claim that the danceability is the same regardless of the explicit contents.
- The danceability distribution of explicit songs is different from the distribution of songs without explicit contents.

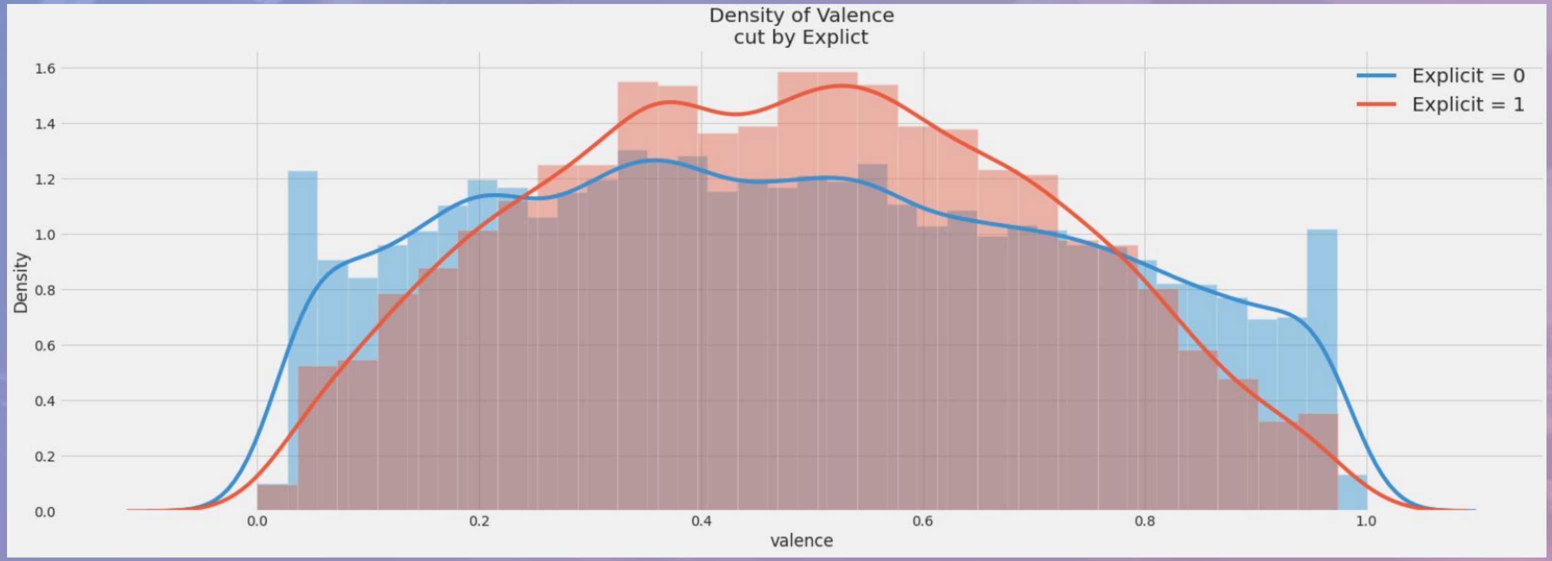
Hypothesis Testing

02

2. Hypothesis test Analysis in valence

(1) Test Setup

- **The Null Hypothesis (H0):** The valence of the tracks without explicit lyrics is the same as those songs with explicit songs.
- **The Alternative Hypothesis (H1):** The valence of the tracks without explicit lyrics is different than those songs with explicit songs.
- The alpha-value will be set to 0.05.



(2) Test Analysis

- The claim is H0.
- Since p-value = 0.559 and it is much higher than the alpha-value = 0.05. We **fail to reject** the null hypothesis.
- The valence distribution of explicit songs is the same as songs without explicit contents.

Hypothesis Testing

03

3. Hypothesis test Analysis in key

(1) Test Setup

- **The Null Hypothesis (H0):** The key of songs is independent with explicit content .
- **The Alternative Hypothesis (H1):** The key of songs is dependent on explicit content.
- The alpha-value will be set to 0.05.

(2) Test Analysis

- The claim is H0.
- Since the p-value is less than the alpha value. We **reject** the null hypothesis.
- The key of songs is dependent on explicit content.

Data Mining Techniques for Explicit Prediction

Use a logistic regression model to predict explicit (0 or 1) using all the other music attributes.

Logistic Regression Model for Explicit Prediction

1. Prediction Setup

- Target attribute - explicit (0 or 1)
- Model evaluation metrics

  (1) classification report

  (2) accuracy

  (3) MSE

  (4) cross validation mean score (cv=10)

2. Prediction Process

- We run a function that automatically selects the best attributes as our training data. After running this function, we get 14 logistic regression models with their evaluation scores. Then, we select the best model based on our model evaluation metrics for explicit prediction.

3. Prediction Results

- The best model we selected using 14 attributes as our training data.
- Model evaluation metrics result:

(1) classification report:

```
(1) Classification_report:
              precision    recall  f1-score   support

           0       0.89      0.98      0.93      7938
           1       0.76      0.37      0.49      1504

    accuracy                           0.88      9442
   macro avg       0.83      0.67      0.71      9442
weighted avg       0.87      0.88      0.86      9442
```

(2) accuracy: 0.88

(3) MSE: 0.119

(4) cross validation mean score (cv=10): 0.879

- This model has high accuracy with low MSE score, which indicates this model performs well in the prediction. Also, this model has a high cross validation mean score, which indicates it has great prediction ability for unseen (test) data.
- One thing we need to be aware of is that the recall of explicit = 1 is 0.37. This indicates this model might not be able to identify the songs which definitely contain explicit content.
- This model needs to be improved for explicit = 1 prediction.

Data Mining Techniques for Popularity Prediction

1. Use different kinds of regression models to predict popularity value (linear regression, decision tree regressor, XGB regressor).
2. Use different kinds of classification models to predict popularity level (Random Forest Classifier, logistic regression (multiclass), and K Nearest Neighbors Classifier).

Data Mining Techniques 01

I. Popularity value prediction using regression models

1. Prediction Setup

- Target attribute - popularity (0-100)
- Model evaluation metric - MSE

2. Prediction Process

- Data pre-processing: Use MinMaxScaler() for year, tempo and duration_mins columns; use OneHotEncoder() for key columns.
- Fit training data for linear regression, decision tree regressor, and XGB regressor model.
- Compare the model evaluation metric.

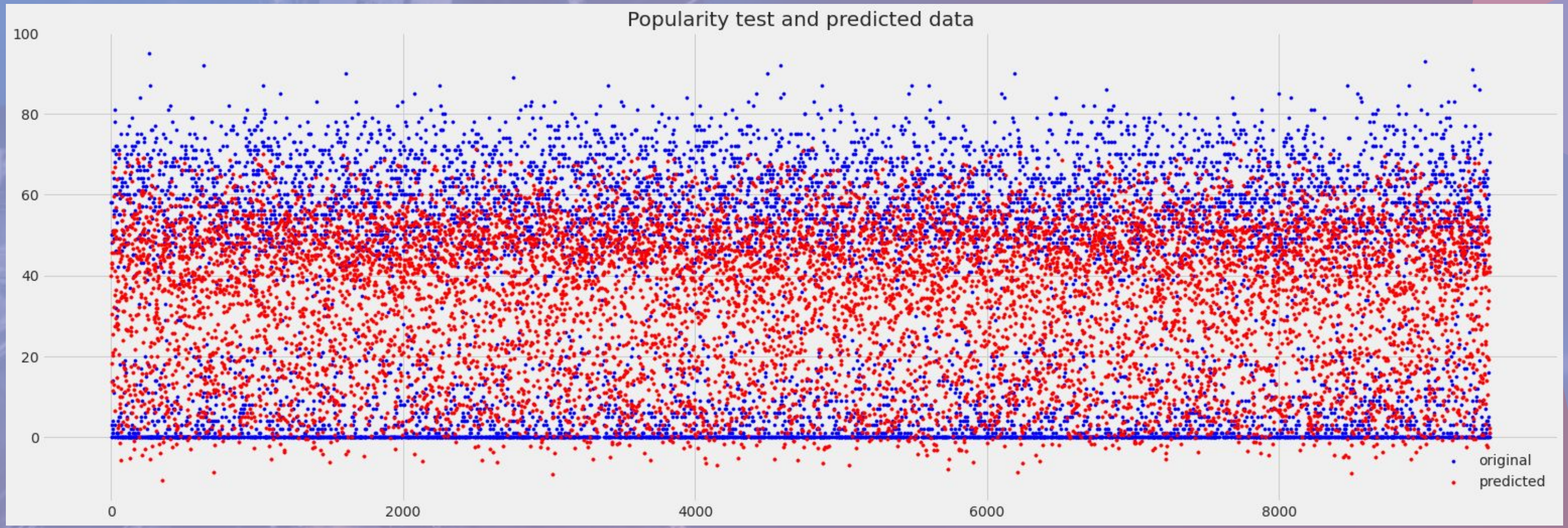
Data Mining Techniques 02

I. Popularity value prediction using regression models

3. Comparison of Regression Models

Model/Metric	Linear Regression	Decision Tree Regressor	XGBRegressor
MSE	558.86	775.46	445.67

- According to our model evaluation metric, we found that the XGB regressor model performs the best for popularity value prediction with the smallest MSE (445.67); while decision tree regressor model performs the worst with the highest MSE (775.46).
- We would like to use xxx model to predict popularity values.
- XGB Regressor Model output visualization of original and predicted popularity values:



Data Mining Techniques 03

II. Popularity level prediction using classification models

1. Prediction Setup

- Target attribute - popularity level (1-4): we divided popularity into four levels using the below criteria:
  Level 1: popularity value between 0 to 25
  Level 2: popularity value between 25 to 50
  Level 3: popularity value between 50 to 75
  Level 4: popularity value between 75 to 100
  Level 4 indicates the song is in the highest popularity group; while level 1 indicates the song is in the lowest popularity group.
- Model evaluation metric
  (1) classification report
  (2) cross validation mean score (cv=6)

2. Prediction Process

- Data pre-processing: Use MinMaxScaler() for year, tempo and duration_mins columns; use OneHotEncoder() for key columns.
- Fit training data for random forest classifier, logistic regression (multi_class), and k nearest neighbors classifier.
- Compare the model evaluation metric.

Data Mining Techniques

04

II. Popularity level prediction using classification models

3. Comparison of Classification Models

Model/Metric	Random Forest Classifier	Logistic Regression (multi_class)	KNN Classifier
Accuracy	0.69	0.52	0.6

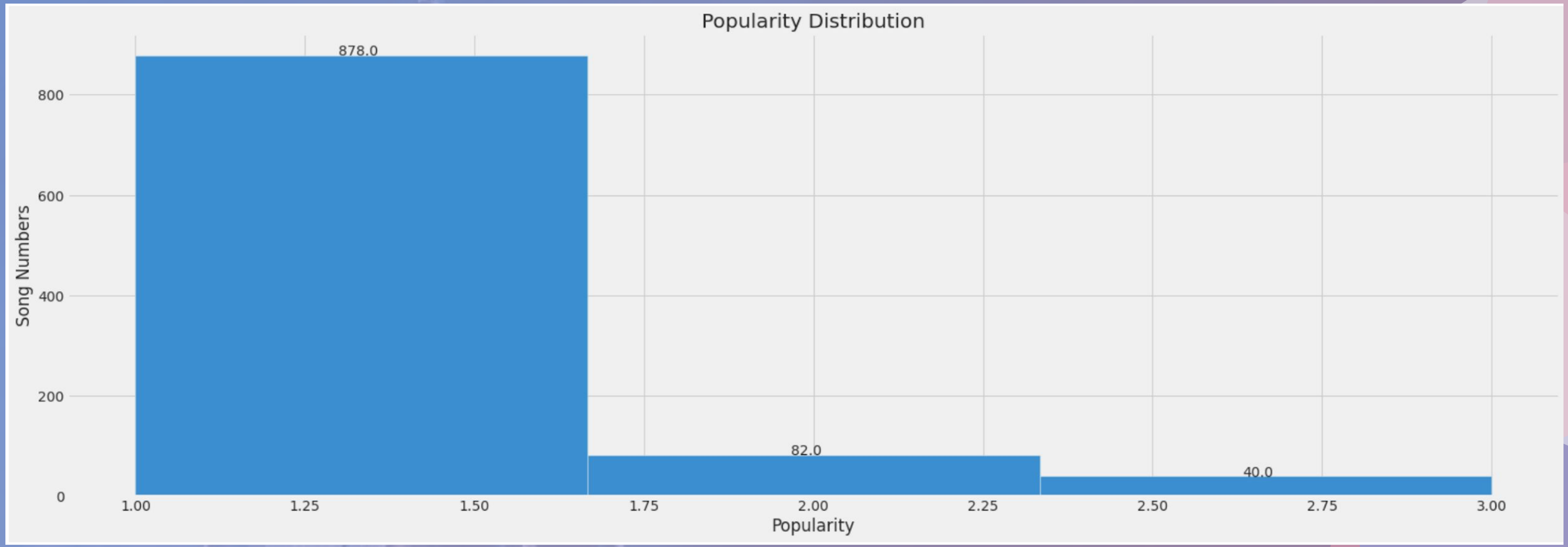
```
Model:  Random Forest
Accuracy on Test Set for Random Forest = 0.69

              precision    recall  f1-score   support

           1       0.78      0.84      0.81      4137
           2       0.50      0.40      0.44      1313
           3       0.66      0.69      0.67      3687
           4       0.32      0.08      0.12       305

    accuracy                           0.69      9442
   macro avg       0.56      0.50      0.51      9442
weighted avg       0.68      0.69      0.68      9442
```

- According to our model evaluation metric, we found that the random forest classifier model performs the best for popularity level prediction with the highest accuracy (0.69); while logistic regression (multi_class) model performs the worst with the lowest accuracy (0.52).
- Random forest classifier is the best model for predicting popularity level. However, according to the classification report, we could find that this model does not perform well in popularity level 2 and level 4. This indicates this model might not be able to correctly identify level 2 and level 4 groups.

Data Mining Techniques 05

II. Popularity level prediction using classification models

4. Predict 1000 Sample Records using Selected Random Forest Model



- We randomly generated 1000 sample records, and used our selected random forest model to predict the popularity level for these 1000 sample data. The plot shows the histogram of predicted popularity level.

Implications

- Gained insights by implementing statistical analysis and visualizations on features.
- Analyzed the trend of music popularity over time.
- Explicit is related to two features - danceability and key. It can be considered when to decide to play explicit music.
- Predicted popularity of songs and forecasted which track can be a hit on the Spotify music platform.
- Predicted whether the song contains explicit content.
- Built machine learning models, regression and classification:
    - XGB regressor performs the best for popularity value prediction with the smallest MSE (445.67).
    - Random forest classifier performs the best for popularity level prediction with the highest accuracy (0.69).
- Prediction model can be applied on music recommendation program for the listeners.
- Musicians, composers, and marketers can take insights of what features make songs popular.

Summary

- In the project, we explored and analyzed the Spotify music dataset from 1920 to 2021. After data cleaning and observation, we deployed exploratory data analysis techniques (EDA) and data visualization. We identified that the popularity of the tracks increased from 1950 to 2000 and decreased after the year 2000.

- Then we used various methodologies to analyze the dataset after creating trained and tested datasets. After running each model, we used model selection techniques to compare the model performance, then decided the optimal models as XGB regressor and Random Forest classifier for the prediction on popularity values and levels.

- Music is composed of various features and it is not easy to disassemble into small pieces. However, through exploring and analyzing the music dataset, we were able to gain some insights and generate applicable models for future use.

Appendix

- **Code Script:**
  [https://colab.research.google.com/drive/1f2HLMbPgPDbxJTvtj568V4P_ZBCqXSMi?usp=sharing&pli=1&authuser=1#scrollTo=U1VBbRxEexwo]

- **References:**

Spotify Dataset 1921-2020, 160k+ Tracks. (2021). Retrieved 22 February 2021, from https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks

Different Classification Techniques - Python. (2021). Retrieved 22 February 2021, from https://www.kaggle.com/coolman/different-classification-techniques-python

Kaggle Notebook - Spotify Time Series Analysis. Retrieved 22 February 2021, from https://www.kaggle.com/richardcsuwandi/spotify-time-series-analysis

Kaggle Notebook - Spotify Popularity Prediction. Retrieved 22 February 2021, from https://www.kaggle.com/anatpeled/spotify-popularity-prediction

Kaggle Notebook - Spotify Usecase : Popularity Prediction. Retrieved 22 February 2021, from https://www.kaggle.com/inesmultrier/spotify-usecase-popularity-prediction

THANK YOU

:)