Data mining project

DBM2, INSA Lyon

- Prediction of music genre -

Zileyah Onafowora
Fiona Eguare
Alyona Teterukovsky

The dataset

- From Kaggle.com *[Link]*
- Extensive data on tracks from Spotify
- 18 columns
- 50 005 rows
- Examples of features are acousticness, mode, energy

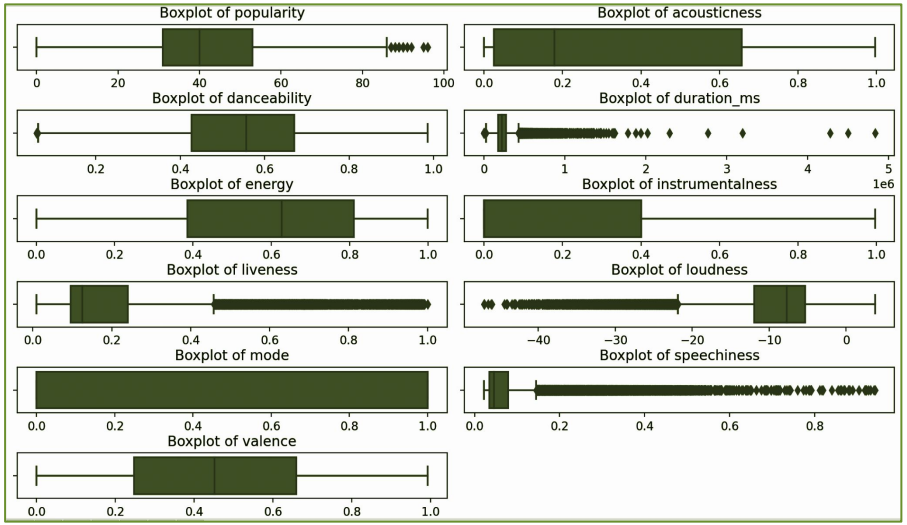
	popularity	acousticness	danceability	duration_ms	energy	instrumentalness	liveness	loudness	mode	speechiness	tempo	valence	music_genre
0	27.0	0.00468	0.652	-1.0	0.941	0.792000	0.1150	-5.201	Minor	0.0748	100.889	0.759	Electronic
1	31.0	0.01270	0.622	218293.0	0.890	0.950000	0.1240	-7.043	Minor	0.0300	115.00200000000001	0.531	Electronic
2	28.0	0.00306	0.620	215613.0	0.755	0.011800	0.5340	-4.617	Major	0.0345	127.994	0.333	Electronic
3	34.0	0.02540	0.774	166875.0	0.700	0.002530	0.1570	-4.498	Major	0.2390	128.014	0.270	Electronic
4	32.0	0.00465	0.638	222369.0	0.587	0.909000	0.1570	-6.266	Major	0.0413	145.036	0.323	Electronic
5	47.0	0.00523	0.755	519468.0	0.731	0.854000	0.2160	-10.517	Minor	0.0412	?	0.614	Electronic

kaggle Spotify®

Table of Contents

Cleaning and pre-processing

- Drop unnecessary columns
- Inconsistencies (?) and missing values (NaN)
- Replace Minor/Major with floats 0.0/1.0
- Normalization/scaling of data
- Outliers - visualize and remove (for classification)



Normalisation (Min-Max [-0.5, 0.5])

	acousticness	danceability	energy	instrumentalness	liveness	loudness	mode	speechiness	valence
0	-0.495301	0.139465	0.441896	0.295181	-0.393642	0.323883	-0.5	-0.442916	0.265121
1	-0.487249	0.107081	0.390804	0.453815	-0.384554	0.287616	-0.5	-0.491628	0.035282
2	-0.496928	0.104922	0.255562	-0.488153	0.029450	0.335381	0.5	-0.486735	-0.164315
3	-0.474498	0.271157	0.200463	-0.497460	-0.351231	0.337724	0.5	-0.264380	-0.227823
4	-0.495331	0.124352	0.087260	0.412651	-0.351231	0.302914	0.5	-0.479341	-0.174395
...	...	...	...	...	...	...	...	...	...
50000	-0.466466	0.421200	0.074237	-0.500000	-0.389602	0.288029	0.5	-0.200228	-0.167339
50001	-0.342369	0.200993	-0.138144	-0.500000	-0.399700	0.233058	0.5	-0.464445	-0.386089
50002	-0.494006	0.183722	0.263576	-0.500000	-0.365368	0.319118	0.5	-0.365500	-0.101815
50003	-0.416566	0.279793	-0.027946	-0.500000	-0.402729	0.327525	-0.5	-0.476297	-0.143145
50004	-0.397590	0.366149	0.142359	-0.500000	-0.235108	0.157492	-0.5	-0.414429	0.271169

50005 rows × 9 columns

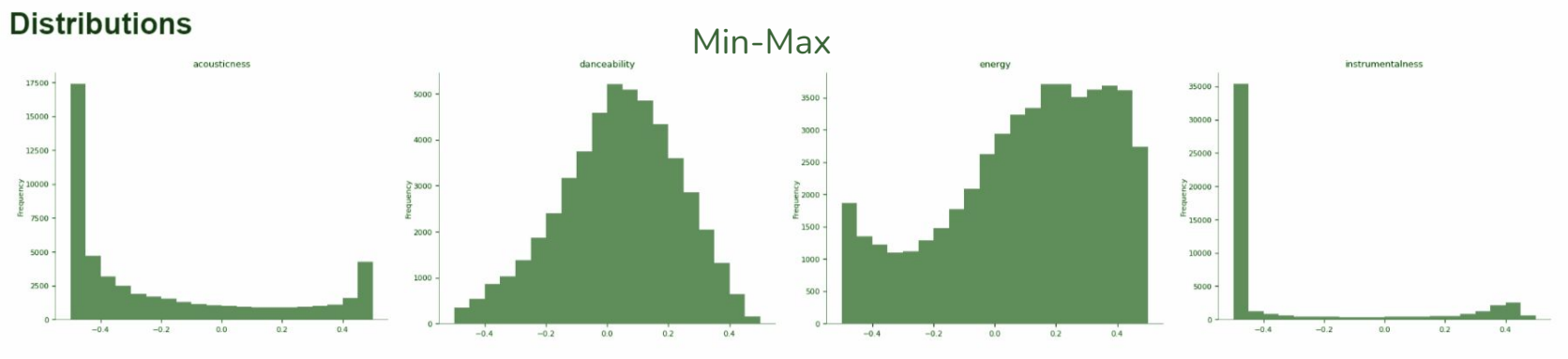
Normalisation (Z-Score)

	acousticness	danceability	energy	instrumentalness	liveness	loudness	mode	speechiness	valence
0	-0.883877	0.524873	1.289863	1.875791	-0.488109	0.638126	-1.339068	-0.185320	1.225061
1	-0.860382	0.356930	1.097090	2.361333	-0.432428	0.339245	-1.339068	-0.627251	0.302428
2	-0.888623	0.345734	0.586807	-0.521807	2.104119	0.732885	0.746773	-0.582861	-0.498807
3	-0.823175	1.207841	0.378914	-0.550294	-0.228267	0.752194	0.746773	1.434438	-0.753745
4	-0.883965	0.446499	-0.048211	2.235338	-0.228267	0.465320	0.746773	-0.515782	-0.539273
...	...	...	...	...	...	...	...	...	...
50000	-0.799738	1.985978	-0.097350	-0.558069	-0.463362	0.342652	0.746773	2.016446	-0.510947
50001	-0.437636	0.843965	-0.898683	-0.558069	-0.525229	-0.110375	0.746773	-0.380638	-1.389067
50002	-0.880098	0.754395	0.617046	-0.558069	-0.314881	0.598859	0.746773	0.517035	-0.247915
50003	-0.654136	1.252626	-0.482897	-0.558069	-0.543789	0.668143	-1.339068	-0.488161	-0.413827
50004	-0.598766	1.700474	0.159681	-0.558069	0.483203	-0.733125	-1.339068	0.073131	1.249341

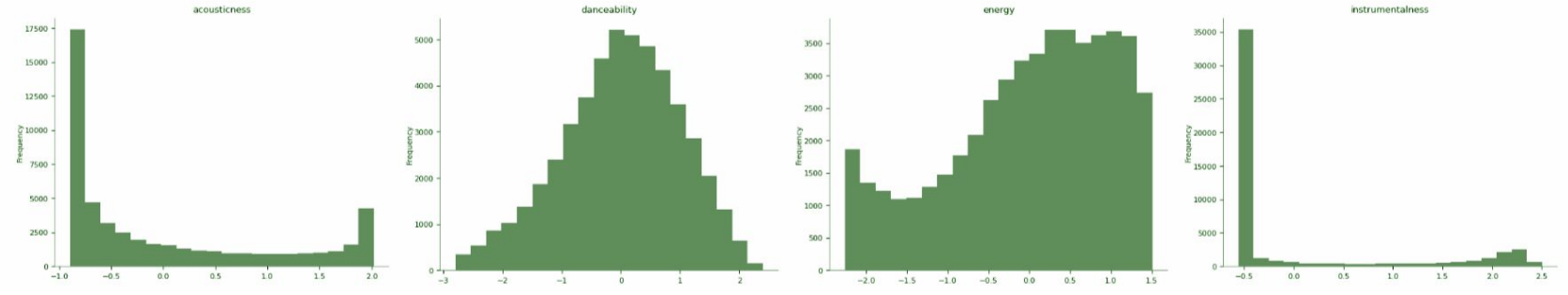
50005 rows × 9 columns

Normalisation (Min-Max Vs. Z-Score)

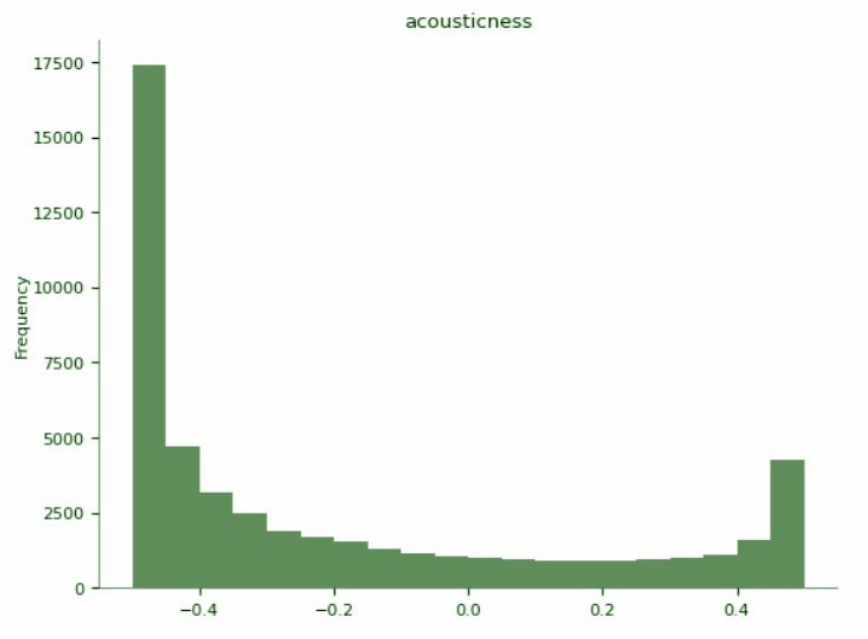
Min-Max



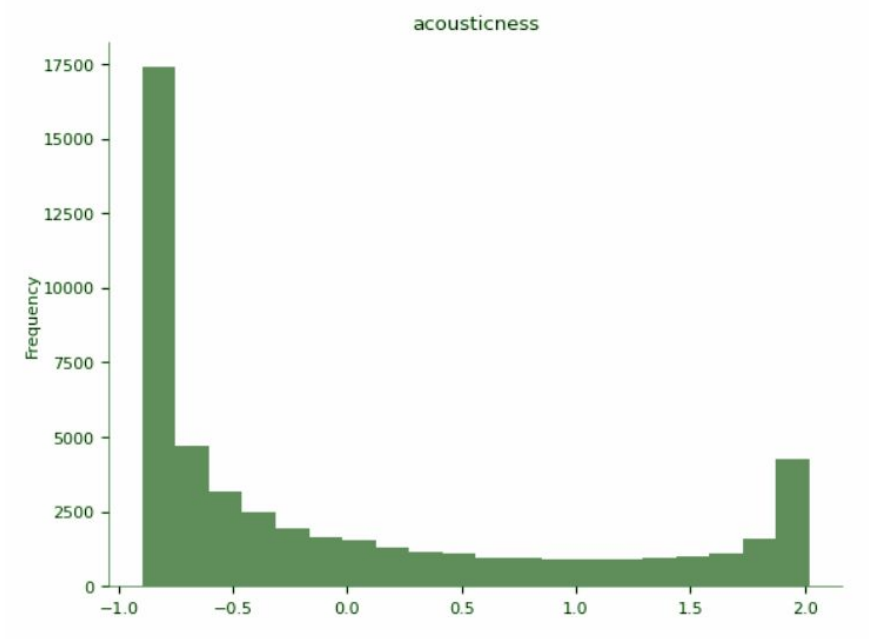
Z-score



Normalisation (Min-Max Vs. Z-Score)



Min-Max



Z-score

Frequent Pattern Mining

1. Clean & preprocess (general)
2. Select the interesting, useable data
3. Discretise the data into high and low
4. Mine the frequent patterns
5. Retrieve “interestingness” data

Selection of Data from Pre-Processed Set

	acousticness	danceability	energy	instrumentalness	liveness	loudness	mode	speechiness	valence
0	0.00468	0.652	0.941	0.79200	0.115	-5.201	0.0	0.0748	0.759
1	0.01270	0.622	0.890	0.95000	0.124	-7.043	0.0	0.0300	0.531
2	0.00306	0.620	0.755	0.01180	0.534	-4.617	1.0	0.0345	0.333
3	0.02540	0.774	0.700	0.00253	0.157	-4.498	1.0	0.2390	0.270
4	0.00465	0.638	0.587	0.90900	0.157	-6.266	1.0	0.0413	0.323
...	...	...	...	...	...	...	...	...	...
50000	0.03340	0.913	0.574	0.00000	0.119	-7.022	1.0	0.2980	0.330
50001	0.15700	0.709	0.362	0.00000	0.109	-9.814	1.0	0.0550	0.113
50002	0.00597	0.693	0.763	0.00000	0.143	-5.443	1.0	0.1460	0.395
50003	0.08310	0.782	0.472	0.00000	0.106	-5.016	0.0	0.0441	0.354
50004	0.10200	0.862	0.642	0.00000	0.272	-13.652	0.0	0.1010	0.765

50005 rows × 9 columns

Discretise (continuous → binary)

	Low acousticness	High acousticness	Low danceability	High danceability	Low energy	High energy	Low instrumentalness	High instrumentalness	Low liveness	High liveness	Low loudness	High loudness	Low mode	High mode	Low speechiness	High speechiness	Low valence	High valence
0	False	True	False	True	False	True	False	True	False	True	True	False	False	True	False	True	False	True
1	False	True	False	True	False	True	False	True	False	True	True	False	False	True	False	True	False	True
2	False	True	False	True	False	True	False	True	False	True	True	False	False	True	False	True	False	True
3	False	True	False	True	False	True	False	True	False	True	True	False	False	True	False	True	False	True
4	False	True	False	True	False	True	False	True	False	True	True	False	False	True	False	True	False	True
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
50000	False	True	False	True	False	True	False	True	False	True	True	False	False	True	False	True	False	True
50001	False	True	False	True	False	True	False	True	False	True	True	False	False	True	False	True	False	True
50002	False	True	False	True	False	True	False	True	False	True	True	False	False	True	False	True	False	True
50003	False	True	False	True	False	True	False	True	False	True	True	False	False	True	False	True	False	True
50004	False	True	False	True	False	True	False	True	False	True	True	False	False	True	False	True	False	True

50005 rows × 18 columns

Frequent Patterns with Apriori

(MM vs Z, min sup = 0.7)

	support	itemsets
0	0.728147	(Low acousticness)
1	0.806999	(Low instrumentalness)
2	0.942786	(Low liveness)
3	0.939426	(High loudness)
4	0.991921	(Low speechiness)
5	0.727627	(High loudness, Low acousticness)
6	0.722008	(Low speechiness, Low acousticness)
7	0.758384	(Low liveness, Low instrumentalness)
8	0.791801	(High loudness, Low instrumentalness)
9	0.799200	(Low speechiness, Low instrumentalness)
10	0.884412	(High loudness, Low liveness)
11	0.935806	(Low speechiness, Low liveness)
12	0.931667	(Low speechiness, High loudness)
13	0.721528	(Low speechiness, High loudness, Low acousticness)
14	0.744686	(High loudness, Low liveness, Low instrumentalness)
15	0.751585	(Low speechiness, Low liveness, Low instrumentalness)
16	0.784222	(Low speechiness, High loudness, Low instrumentalness)
17	0.877592	(Low speechiness, High loudness, Low liveness)
18	0.738046	(Low speechiness, High loudness, Low liveness, Low instrumentalness)

	support	itemsets
0	0.756184	(Low instrumentalness)
1	0.738866	(Low speechiness)

Frequent Patterns with Apriori
(min sup = 0.8, MM)

	support	itemsets
0	0.806999	(Low instrumentalness)
1	0.942786	(Low liveness)
2	0.939426	(High loudness)
3	0.991921	(Low speechiness)
4	0.884412	(High loudness, Low liveness)
5	0.935806	(Low speechiness, Low liveness)
6	0.931667	(Low speechiness, High loudness)
7	0.877592	(Low speechiness, High loudness, Low liveness)

“Interestingness” of frequent patterns
(min conf = 0.7, MM)

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
0	(High loudness)	(Low liveness)	0.939426	0.942786	0.884412	0.941438	0.998571
1	(Low liveness)	(High loudness)	0.942786	0.939426	0.884412	0.938083	0.998571
2	(Low speechiness)	(Low liveness)	0.991921	0.942786	0.935806	0.943429	1.000682
3	(Low liveness)	(Low speechiness)	0.942786	0.991921	0.935806	0.992597	1.000682
4	(Low speechiness)	(High loudness)	0.991921	0.939426	0.931667	0.939255	0.999818
5	(High loudness)	(Low speechiness)	0.939426	0.991921	0.931667	0.991740	0.999818
6	(Low speechiness, High loudness)	(Low liveness)	0.931667	0.942786	0.877592	0.941959	0.999123
7	(Low speechiness, Low liveness)	(High loudness)	0.935806	0.939426	0.877592	0.937792	0.998261
8	(High loudness, Low liveness)	(Low speechiness)	0.884412	0.991921	0.877592	0.992289	1.000372
9	(Low speechiness)	(High loudness, Low liveness)	0.991921	0.884412	0.877592	0.884740	1.000372
10	(High loudness)	(Low speechiness, Low liveness)	0.939426	0.935806	0.877592	0.934179	0.998261
11	(Low liveness)	(Low speechiness, High loudness)	0.942786	0.931667	0.877592	0.930850	0.999123

Frequent Patterns with Apriori
(min sup = 0.5, Z)

	support	itemsets
0	0.632577	(Low acousticness)
1	0.521428	(High danceability)
2	0.558484	(High energy)
3	0.756184	(Low instrumentalness)
4	0.681932	(Low liveness)
5	0.652175	(High loudness)
6	0.641916	(High mode)
7	0.738866	(Low speechiness)
8	0.510449	(Low valence)
9	0.530167	(Low instrumentalness, Low acousticness)
10	0.534947	(High loudness, Low acousticness)
11	0.512269	(High energy, High loudness)
12	0.570983	(High loudness, Low instrumentalness)
13	0.519488	(Low speechiness, Low instrumentalness)
14	0.526507	(Low speechiness, Low liveness)

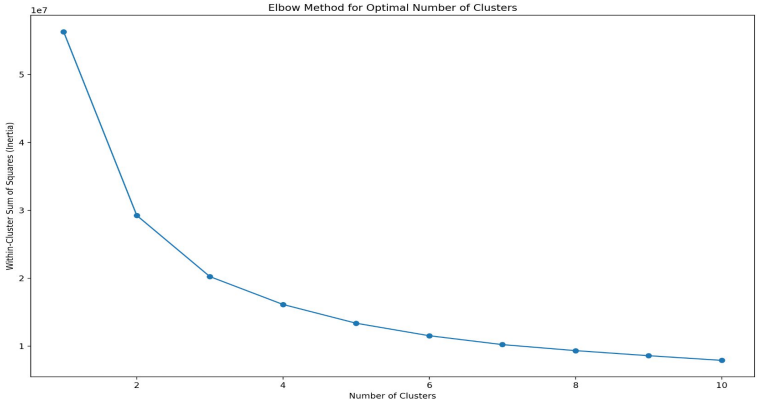
“Interestingness” of frequent patterns
(min conf = 0.7, Z)

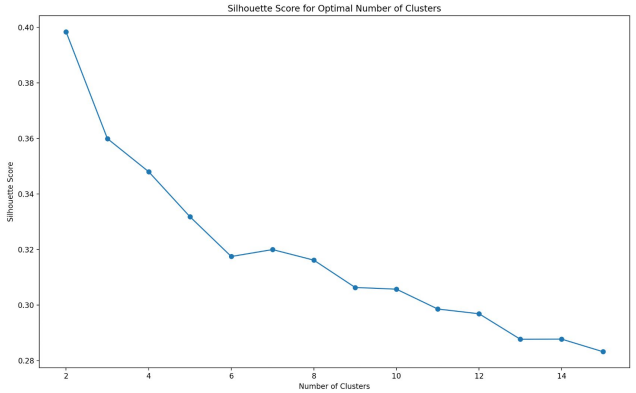
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
0	(Low acousticness)	(Low instrumentalness)	0.632577	0.756184	0.530167	0.838107	1.108337
1	(High loudness)	(Low acousticness)	0.652175	0.632577	0.534947	0.820250	1.296681
2	(Low acousticness)	(High loudness)	0.632577	0.652175	0.534947	0.845663	1.296681
3	(High energy)	(High loudness)	0.558484	0.652175	0.512269	0.917249	1.406446
4	(High loudness)	(Low instrumentalness)	0.652175	0.756184	0.570983	0.875506	1.157794

Clustering

Algorithm Used

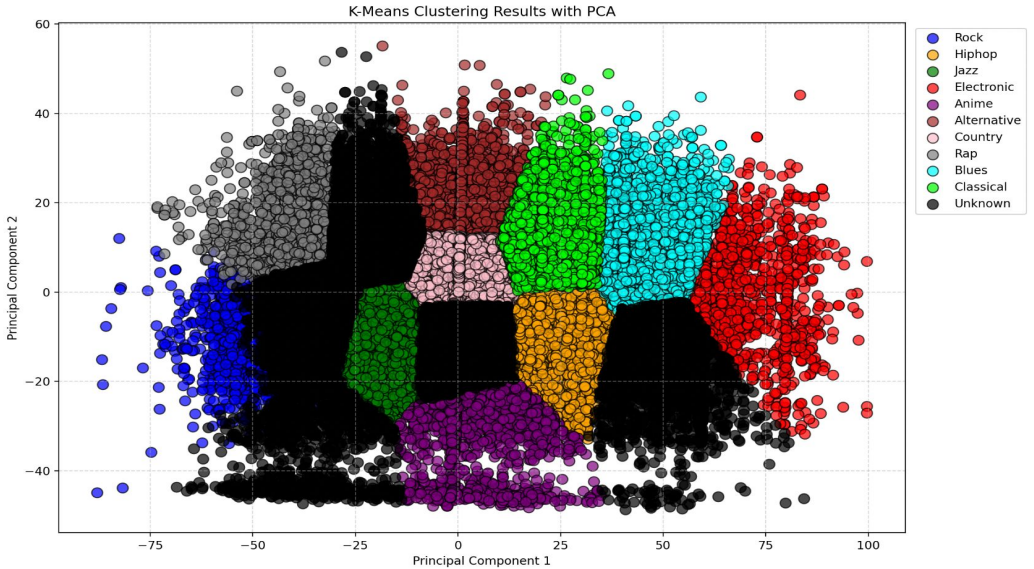
- K Means clustering algorithm to group similar music genres
- Used Key Features such as: energy, danceability, mode, and more
- Though having established clusters
- Elbow Method and Silhouette Method to find Optimal Number of Clusters

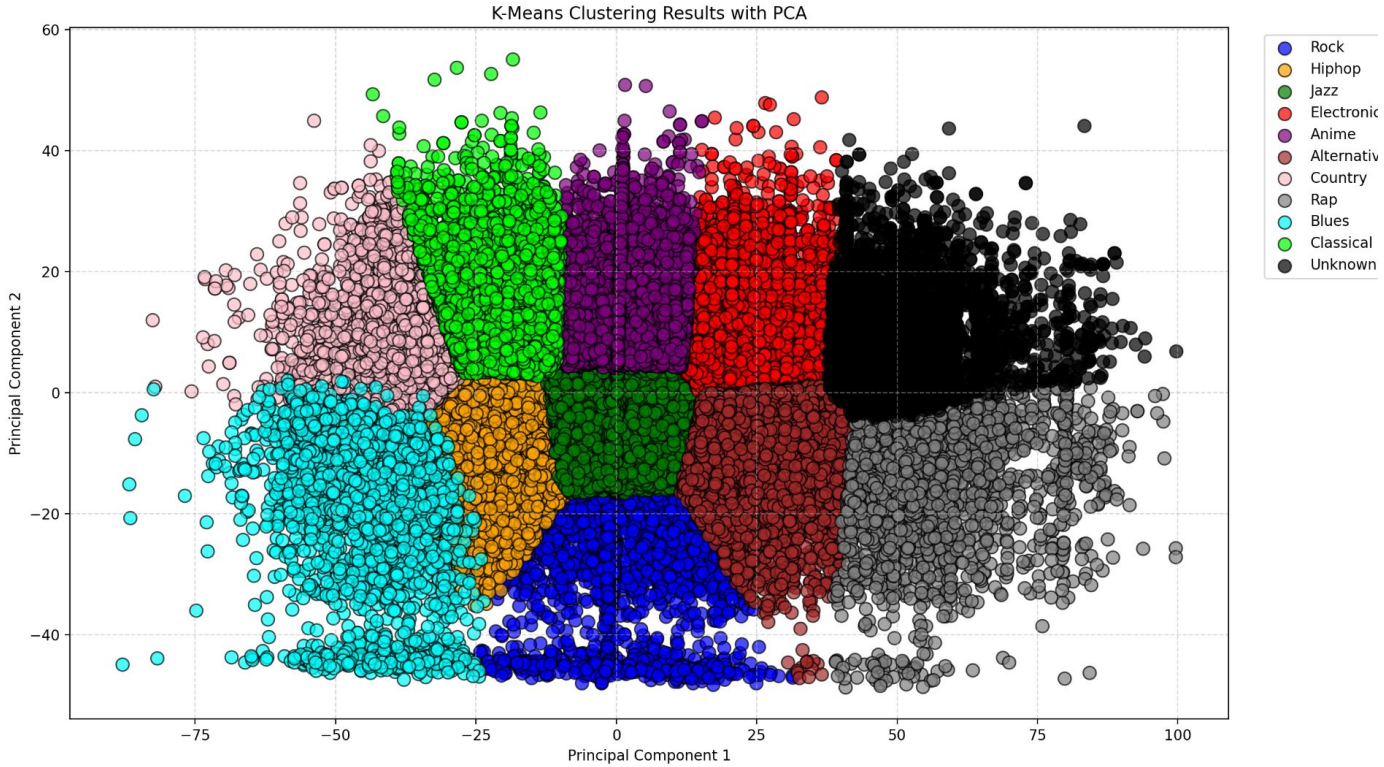




Clustering Visualization

- Apply K Means to Key Features
- Reintroduced Valued featured Genre.
- Label Encoding to represent music with numeric labels
- Mapping cluster back to respective genres
- Principal Component Analysis(PCA)
    - Dimension Reduction. Challenging to visualize without it because of the numerous features




K-Means Clustering Results with PCA
Principal Component 2
Principal Component 1
60
40
20
0
−20
−40
−75
−50
−25
0
25
50
75
100
Rock
Hiphop
Jazz
Electronic
Anime
Alternativ
Country
Rap
Blues
Classical
Unknown

Classification

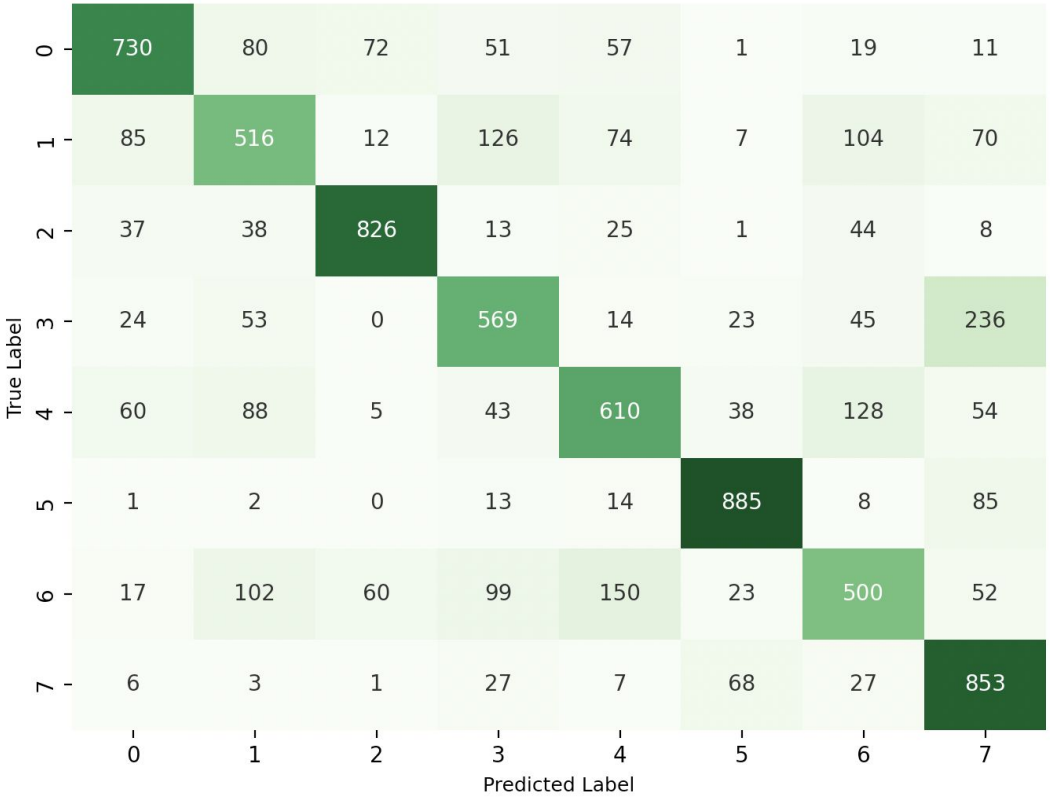
Random Forest Classifier

- Started with decision tree (low accuracy)
- sklearn library
- Split data: 80% train and 20% validation
- n_estimators = 100

- Check for class imbalance
- No need for class weightning

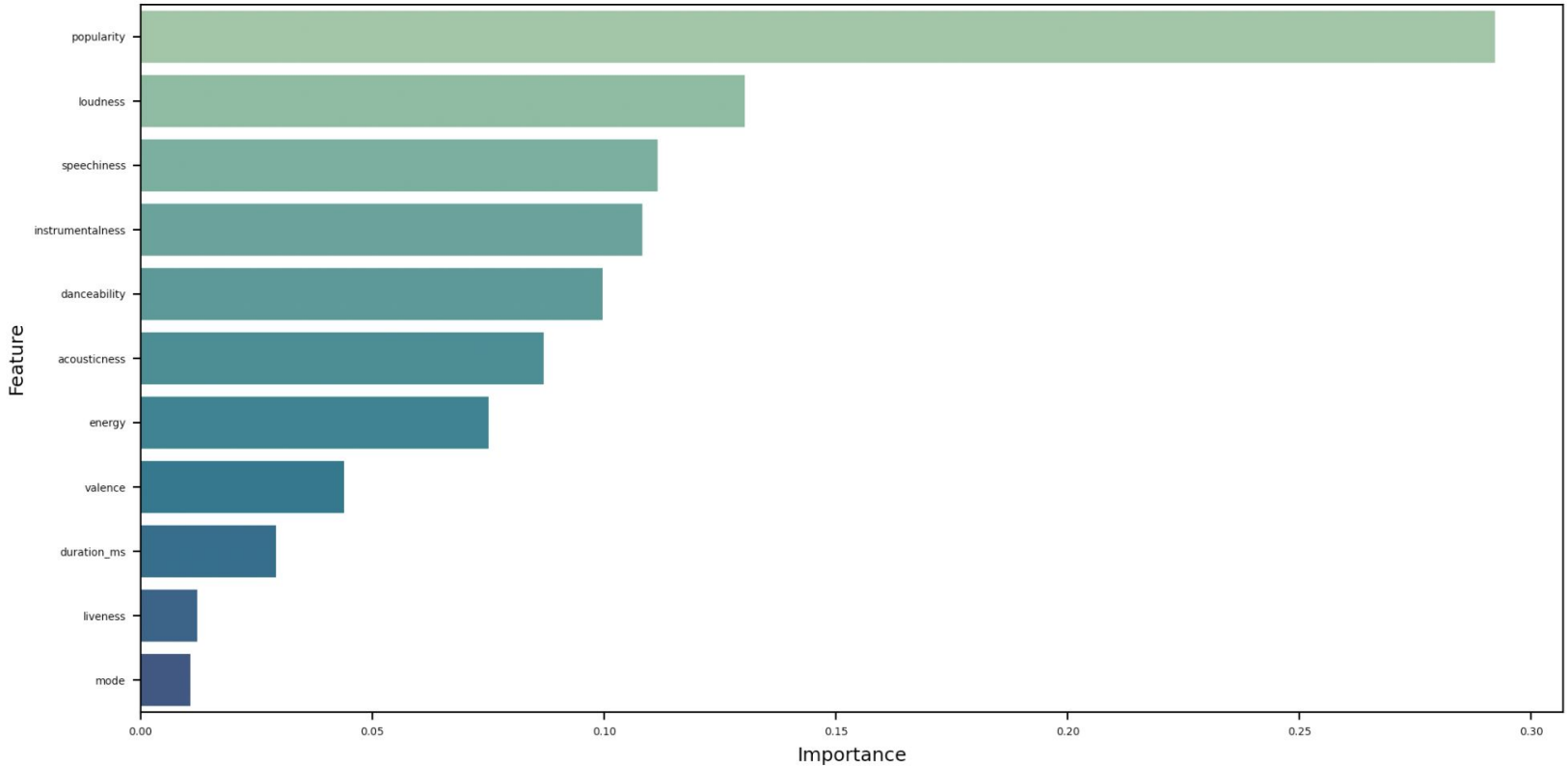
Classification

Confusion Matrix:

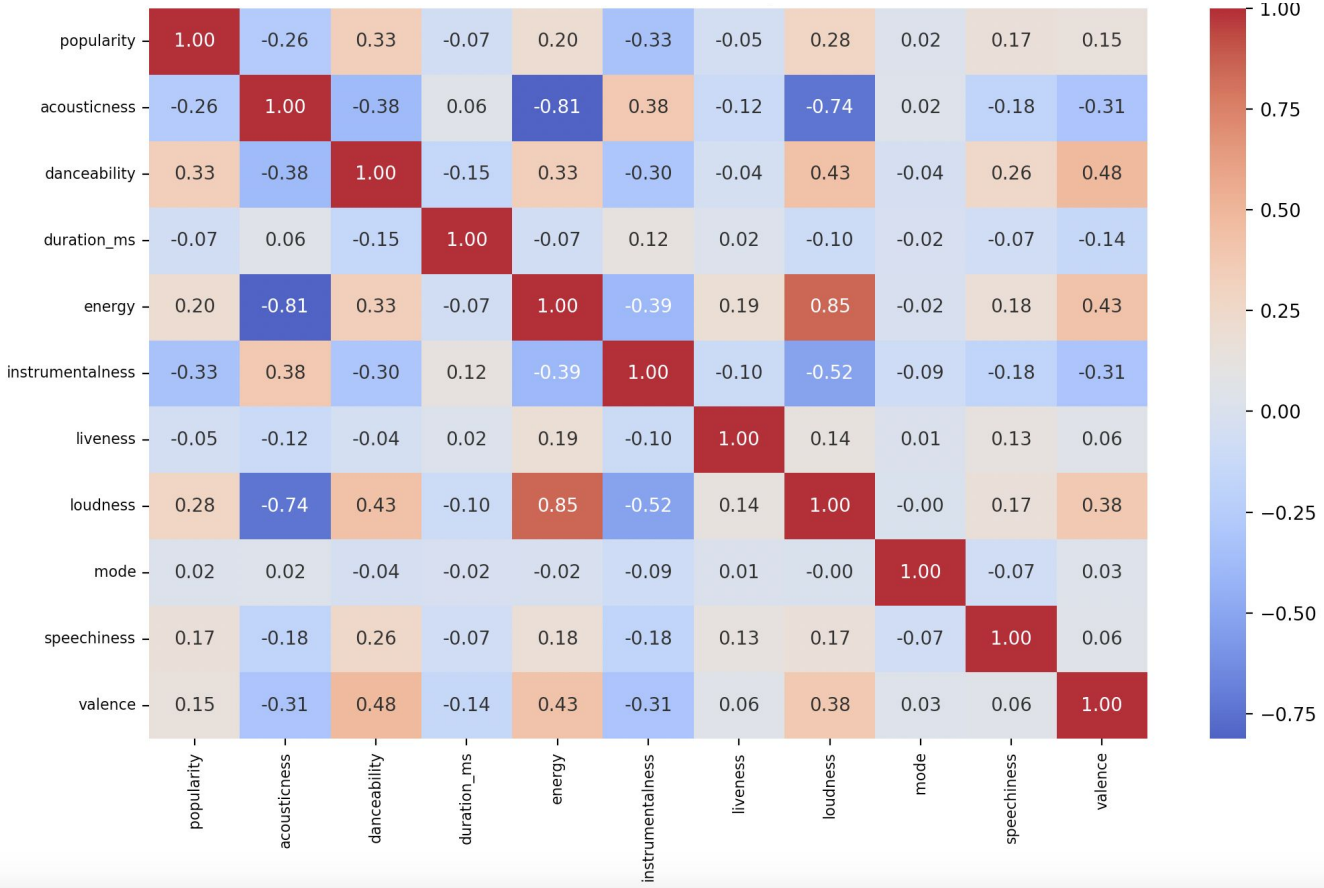


Classification

Feature Importance:



Correlation Heatmap:



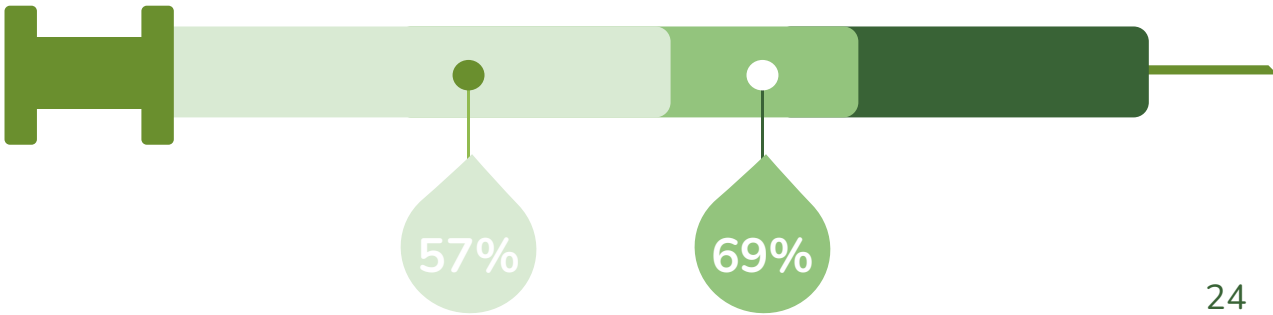
Classification - Results Analysis

Accuracy - 69%

- Mean 0.57 before removing rows Alternative and Rap
- Some genres have better accuracy than others
- Nature of our data

Future improvements:

- Feature Engineering
  - New features
- Try with other algorithms:
  - Ensemble methods, boosting



Evaluation

Pattern mining

- Usefulness in terms of genre prediction could be improved by discretising into more categories
- Methods produce same values
- Methods could be evaluated on time efficiency

Clustering

- Lots of unknown values

Classification

- Circa 70% accuracy
- Generate new features
- Ensemble methods

Thank you!

Questions?